สัญญาเลขที่ TRG5180005

รายงานวิจัยฉบับสมบูรณ์

โครงการ การหาสนิปซ์ที่สำคัญสำหรับการจำแนกประชากรโดยวิธีแสปซย่อย

อภิชาติ อินทรพานิชย์
ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย

บทคัดย่อ

รหัสโครงการ TRG5180005

ชื่อโครงการ การหาสนิปซ์ที่สำคัญสำหรับการจำแนกประชากรโดยวิธีแสปซย่อย

ชื่อนักวิจัย อภิชาติ อินทรพานิชย์ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

E-mail address: apichart.intarapanich@nectec.or.th

ระยะเวลาโครงการ 1 ปี

ปัจจุบันข้อมูลของจีโนไทป์มีขนาดใหญ่มากขึ้น ทำให้การอนุมานจำนวนกลุ่มประชากรย่อยและการจำแนกคนเข้าในกลุ่มประชากรย่อยเป็นเรื่องยาก วิธีการที่เป็นที่นิยมและมีประสิทธิภาพในการคำนวนข้อมูลที่มีขนาดใหญ่คือ Principal Components Analysis (PCA) ในปัจจุบันนี้ PCA สามารถตรวจจับหาโครงสร้างของประชากรได้เท่านั้นไม่สามารถที่จะหาจำนวนกลุ่มของประชากรย่อยได้ ดังนั้นคณะวิจัยจึงทำการพัฒนาระเบียบวิธีโดยปรับเปลี่ยนเทคนิคของ PCA เพื่อทำการหาจำนวนกลุ่มประชากรย่อยและสามารถตรวจสอบหากลุ่มประชากรที่เหมาะสมกับบุคคลได้ ระเบียบวิธีที่ทำการพัฒนาขึ้นนี้เรียกว่า iterative pruning PCA (ipPCA) สามารถใช้กับข้อมูลจีโนไทป์มีขนาดใหญ่ได้อย่างมีประสิทธิภาพ โดยระเบียบวิธี ipPCA ได้ถูกทดสอบประสิทธิภาพกับข้อมูลจำลอง ผลการทดลองแสดงให้เห็นว่าสำหรับข้อมูลกลุ่มประชากรที่ไม่ซับซ้อนผลที่ได้สอดคล้องกับระเบียบวิธี STRUCTURE และ AWclust อย่างไรก็ตาม สำหรับข้อมูลของกลุ่มประชากรที่ซับซ้อน มีเพียงระเบียบวิธี ipPCA เท่านั้นที่สามารถจำแนกกลุ่มประชากรได้อย่างถูกต้อง นอกจากนี้ระเบียบวิธี ipPCA ถูกทดสอบกับข้อมูลจริงที่มีขนาดของข้อมูลจีโนไทป์ใหญ่มากซึ่งไม่สามารถวิเคราะห์ได้โดยระเบียบวิธี STRUCTURE และ AWclust

Abstract

Project Code : TRG5180005

Project Title :

Investigator : Apichart.intarapanich National Electronics and Computer Technology Center

E-mail address: apichart.intarapanich@nectec.or.th

Project Period : 1 year

As genotypic datasets become ever larger, it is increasingly difficult to correctly estimate the number of subpopulations and assigning individuals to them. The computationally efficient non-parametric, chiefly Principal Components Analysis (PCA)-based methods are thus becoming increasingly relied upon for population structure analysis. Current PCA-based methods can accurately detect structure; however, the accuracy in resolving subpopulations and assigning individuals to them is wanting. When subpopulations are closely related to one another, they overlap in PCA space and appear as a conglomerate. This problem is exacerbated when some subpopulations in the dataset are genetically far removed from others. We propose a novel PCA-based framework which addresses this shortcoming.

A novel population clustering program called iterative pruning PCA (ipPCA) was developed which assigns individuals to subpopulations and infers the total number of subpopulations present. Genotypic data from simulated and real population datasets with different degrees of structure were analyzed by the proposed method. For datasets with simple structures, the subpopulation assignments of individuals made by ipPCA were consistent with the STRUCTURE and AWclust algorithms. On the other hand, highly structured populations containing many closely related subpopulations could be accurately resolved only by ipPCA, and not by other methods. The algorithm is computationally efficient and not constrained by the dataset complexity. This systematic clustering approach removes the need for prior population labels, which could be advantageous when cryptic stratification is encountered in datasets of individuals otherwise assumed to be homogenous.

สารบัญ

บทที่ 1 บทนำ

ความถี่ของอัลลีลในแต่ละประชากรนั้นมีความแตกต่างกันเนื่องจากกลไกของวิวัฒนาการ เช่น การอพยพย้านถิ่นฐาน การคัดเลือกโดยธรรมชาติ ภัยพิบัติและโรคระบาด ดังนั้นกลุ่มตัวอย่างในประชากรจึงสามารถถูกปรับเปลี่ยนเป็นโครงสร้างประชากรย่อย การศึกษาวิวัฒนาการของประชากรใดๆ จำเป็นจะต้องอาศัยระเบียบวิธีที่มีประสิทธิภาพพอเพียงเพื่อการบ่งชี้การมีอยู่ของโครงสร้างประชากรย่อย และยิ่งไปกว่านั้นโครงสร้างประชากรย่อยยังเป็นปัจจัยหลักสำหรับการศึกษาระบาดวิทยาเชิงพันธุกรรมที่สามารถสร้างความคลุมเครือในการวิเคราะห์หาการเกี่ยวเนื่องระหว่างยีนและโรคได้ [1] การวิเคราะห์หาโครงสร้างประชากรโดยทั่วไปแล้วประกอบด้วย 4 เป้าหมายหลักคือ 1) การตรวจหาโครงสร้างประชากร 2) การกำหนดกลุ่มประชากรให้กับตัวอย่าง 3) หาจำนวนกลุ่มของประชากรย่อย และ 4) สามารถหาสัดส่วนของประชากรในแต่ละกลุ่มประชากร [2] จากเทคโนโลยีของการจีโนไทป์ในปัจจุบันที่สามารถทำได้ทีละหลายๆ สนิปพร้อมๆ กันถึงครั้งละห้าแสนสนิปในหนึ่งแผ่นอาร์เรย์ ทำให้ข้อมูลจีโนไทป์มีปริมาณเพิ่มขึ้นมหาศาลในเวลาอันรวดเร็ว ยิ่งเพิ่มความยากลำบากในวิเคราะห์ข้อมูลเพื่อตอบคำถามด้านพันธุศาสตร์ประชากร รวมไปถึงระบาดวิทยาเชิงพันธุศาสตร์มากขึ้น ดังนั้นการพัฒนาระเบียบวิธีที่มีอยู่แล้วรวมไปถึงการคิดค้นระเบียบวิธีใหม่ๆ ที่มีประสิทธิภาพมากกว่าเดิมเพื่อใช้ในการค้นหาโครงสร้างประชากรจึงเป็นเรื่องที่จำเป็นอย่างยิ่ง หลายๆ ระเบียบวิธีในการคำนวณเพื่อค้นหาโครงสร้างประชากรได้ถูกเสนอขึ้น โดยทั่วไปแล้วแบ่งเป็นสองแบบคืออ้างอิงโมเดลและแบบไม่อ้างอิงโมเดล วิธีการแบบอ้างอิงโมเดลนั้นถือว่าเป็นวิธีที่ได้รับความนิยมสูงเนื่องจากเป็นเทคนิคทางด้านสถิติโดยอาศัยความถี่ของอัลลีลเพื่อการตรวจสอบโครงสร้างและกำหนดกลุ่มของประชากรให้กับแต่ละตัวอย่าง วิธีนี้มีความแม่นยำสูงแต่อย่างไรก็ตามการทำนายจำนวนแท้จริงของประชากรย่อย (K) นั้นยังไม่เที่ยงตรงมากนัก เนื่องจากเป็นวิธีอนุมานค่าจากสถิติจึงต้องมีการคำนวณหลายๆ ครั้งแล้วนำผลที่ได้บ่อยครั้งที่สุดเป็นคำตอบ และยังอ้างอิงข้อสรุปของฮาร์ดี-ไวน์เบิร์กซึ่งอาจไม่เป็นจริงตามนั้นในหลายๆ ข้อมูล ยกตัวอย่างเช่น การสุ่มตัวอย่างข้อมูลแบบมีอคติเนื่องจากข้อจำกัดของการทำการทดลอง ยิ่งไปกว่านั้นวิธีการแบบอ้างอิงโมเดลยังใช้ทรัพยากรเวลาในการคำนวณสูงเมื่อเทียบกับแบบไม่อ้างอิงโมเดล

สำหรับการค้นหาโครงสร้างประชากรด้วยวิธีการแบบไม่อ้างอิงโมเดลที่นิยมใช้มากที่สุดคือการประยุกต์เอาการวิเคราะห์ของไอเกนมาใช้ ยกตัวอย่างเช่น principal component analysis (PCA) และวิธีการที่เกี่ยวเนื่องกันที่ชื่อว่า singular value decomposition (SVD) โดยทั่วไปแล้ววิธีการนี้สามารถคำนวณกับข้อมูลขนาดใหญ่โดยใช้เวลาไม่มากนัก แต่ PCA เพียงอย่างเดียวก็ยังไม่สามารถที่จะประมาณค่า K ได้ ดังนั้นระเบียบวิธีการจัดกลุ่มจึงถูกนำมาใช้ในการจำแนกตัวอย่างสู่ประชากรย่อย ข้อมูลจะถูกเชื่อมโยงความสัมพันธ์ตามระยะห่าง

ของพันธุกรรมโดยอาศัยการทำงานของ PCA และด้วยจำนวนแกนสำคัญ (principal component) เพียงไม่มาก PCA จะสามารถแสดงตัวอย่างให้อยู่ตามกลุ่มของประชากรย่อย โดยถ้าประชากรย่อยที่ห่างกันมากบนแกนสำคัญจะแสดงถึงการมีระยะห่างของพันธุกรรมมาก ในทางตรงกันข้ามถ้าประชากรย่อยที่ใกล้กันบนแกนสำคัญจะแสดงถึงการมีระยะห่างของพันธุกรรมน้อยตามไปด้วย ในกลุ่มของประชากรย่อยที่ใกล้กันมากๆ การที่จะชี้ชัดให้เห็นถึงความแตกต่างและแยกย่อยออกไปได้นั้นจะต้องอาศัยจำนวนแกนสำคัญจำนวนมากกว่าเดิมซึ่งสามารถหาจำนวนได้โดยมีการทดลองมาก่อน วิธีการจัดกลุ่มจะถูกประยุกต์ใช้ด้วยจำนวนแกนสำคัญที่ต้องเหมาะสมต่อข้อมูลเท่านั้น เนื่องจากการเพิ่มจำนวนแกนมากเกินไปอาจเป็นการรบกวนการวิเคราะห์ข้อมูลโครงสร้างด้วยข้อมูลที่ไม่มีความสำคัญอื่นๆ ซึ่งเราทราบดีอยู่แล้วว่าข้อมูลโครงสร้างประชากรที่สำคัญนั้นมีจำนวนน้อยกว่าข้อมูลที่ไม่มีความสำคัญมาก ดังนั้นการใช้แกนที่สำคัญจำนวนมากๆ กับวิธีการจัดกลุ่มนั้นไม่สามารถเพิ่มความแม่นยำของการค้นหาโครงสร้างประชากรแต่อย่างใด

วิธีหนึ่งซึ่งสามารถเพิ่มประสิทธิภาพในการแยกย่อยกลุ่มประชากรที่ใกล้ชิดกันได้ถูกต้องยิ่งขึ้นคือการกำจัดประชากรย่อยที่มีระยะห่างทางพันธุกรรมมากที่สุดออกก่อน จากนั้นจึงประยุกต์ PCA บนข้อมูลที่เหลือ วิธีการนี้สามารถช่วยทำให้การแยกย่อยกลุ่มประชากรที่ใกล้ชิดที่เหลืออยู่ได่ดีขึ้น วิธีการตัดออกเช่นนี้ถือว่าเป็นประโยชน์กับการวิเคราะห์ในระดับที่ต้องการความละเอียดสูง อย่างไรก็ตามวิธีนี้ยังจำเป็นต้องใช้ข้อมูลฉลากเพื่อบอกว่าตัวอย่างใดควรตัดออกก่อนบ้าง วิธีนี้จึงไม่สามารถใช้ได้ในกรณีที่ข้อมูลมีฉลากที่เหมือนกันได้ ยกตัวอย่างเช่น การศึกษาความเกี่ยวเนื่องกันของยีนและโรค โดยมีตัวแปรควบคุมคือเชื้อชาติและภูมิภาคที่กำเนิด
ทั้งๆ ข้อดีของวิธีการหาโครงสร้างประชากรแบบไม่อ้างอิงโมเดลมีมากมาย แต่ก็ไม่มีวิธีใดที่สามารถที่จะกำหนดประชากรย่อยให้กับตัวอย่างได้อย่างสมบูรณ์และเป็นระบบ สิ่งที่ตามมาคือวิธีเหล่านี้ก็ยังมีความสามารถไม่เพียงพอที่จะหาจำนวนประชากร K ได้อย่างถูกต้องนั่นเอง

ในงานวิจัยนี้นำเสนอระเบียบวิธีที่ไม่อ้างอิงโมเดลที่เรียกว่า iterative pruning PCA (ipPCA) ซึ่งสามารถตรวจสอบการมีอยู่ของโครงสร้างประชากร การกำหนดกลุ่มของประชากรย่อยให้ตัวอย่าง การทำนายจำนวนของประชากรย่อย (K) ด้วยความแม่นยำมากกว่าวิธีอื่นๆ ipPCA ใช้วิธีแบบฮิวริสติคเพื่อการค้นหาประชากรย่อยในระดับที่เหมาะสมกับความเป็นจริงของข้อมูลจีโนไทป์ด้วยระเบียบวิธีแบบทำซ้ำจนหยุดกระบวนการ

บทที่ 2 ระเบียบวิธี

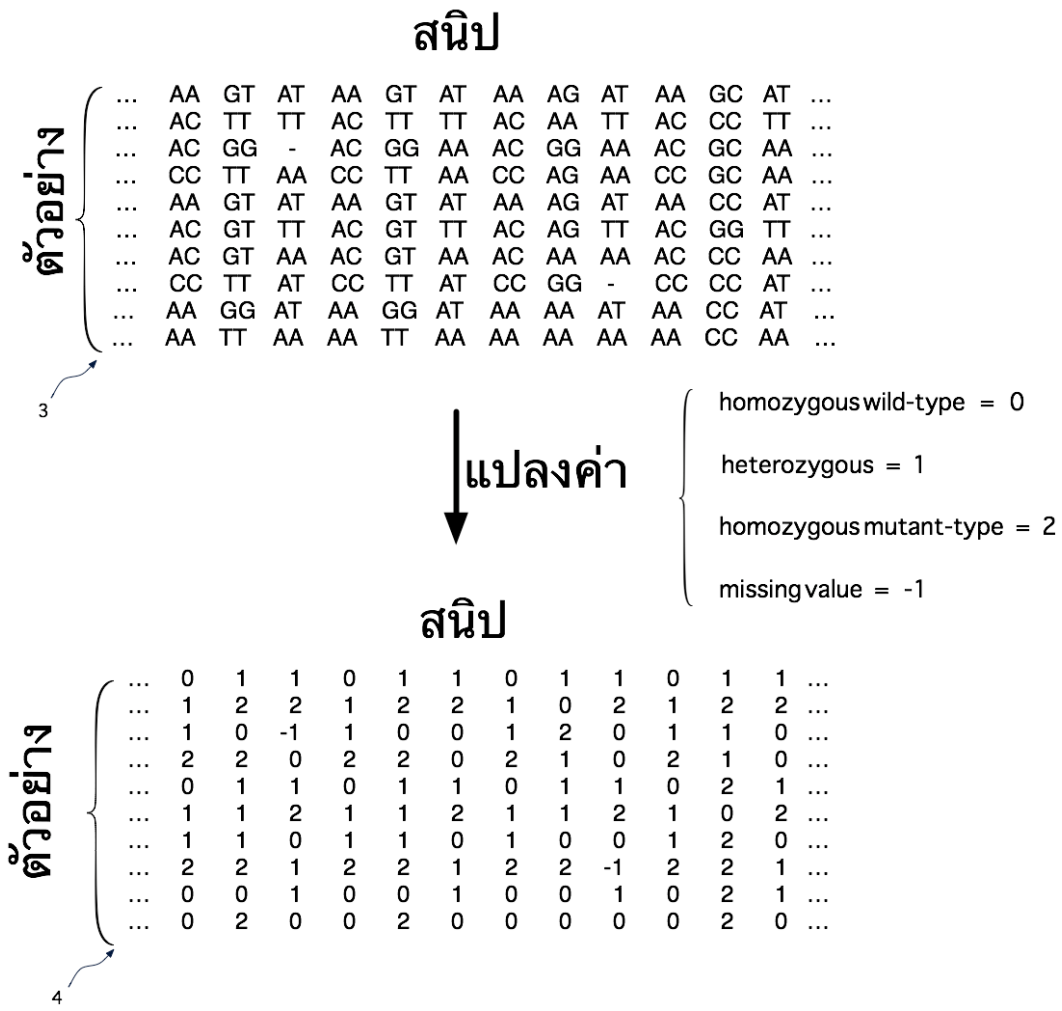
ในงานวิจัยนี้ได้ทำการพัฒนาระเบียบวิธีเพื่อทำการหาจำนวนกลุ่มประชากรย่อยจากข้อมูลของกลุ่มประชากรที่มีโดยที่ไม่จำเป็นต้องทราบจำนอนของกลุ่มประชากรย่อยล่วงหน้า นอกเหนือจากนี้ระเบียบวิธีนี้ยังสามารถระบุบุคคลไปอยู่ในกลุ่มประชากรย่อยที่เหมาะสมได้ ระเบียบวิธีที่ทำการพัฒนาขึ้นนี้ เป็นแบบที่ไม่ขึ้นกับข้อสมติฐานของกลุ่มประชากร โดยการพิจารณาว่ากลุ่มประชากรย่อยที่เหมาะสมนี้จะดูจากความคล้ายคลึงกันทางพันธุกรรม ซึ่งกลุ่มประชากรย่อยเดียวกันจะมีความคล้ายคลึงกันทางพันธุกรรมมากที่สุด

ระเบียบวิธีที่พัฒนาขึ้นนี้นี้ใช้ขั้นตอนเดียวกับ PCA ซึ่งได้มีการนำมาใช้กับการวิเคราะห์กลุ่มประชากร แต่ไม่สามารถจำแนกหากลุ่มประชากรย่อยได้ละเอียด งานวิจัยนี้ได้ทำการพัฒนาปรับปรุงขั้นตอนของ PCA เพื่อให้สามารถหาจำนวนกลุ่มประชากรย่อยได้อย่างถูกต้องแม่นยำมากขึ้น โดยระเบียบวิธีใหม่นี้ได้ทำการแยกกลุ่มประชากรย่อยออกไปในแต่ละขั้นที่ทำ PCA ซึ่งจะสามารถทำให้ในขั้นตอนต่อไปของ PCA จะสามารถเห็นกลุ่มประชากรย่อยที่ชัดเจนมากขึ้น ซึ่งวิธีที่พัฒนาขึ้นมานี้เรียกว่า Iterative Pruning PCA หรือ ipPCA เนื่องจากระเบียบวิธี ipPCA เป็นขั้นตอนที่ต้องทำซ้ำไปเรื่อย ๆ ดังนั้นจึงจำเป็นต้องมีการทดสอบว่าจะหยุดการทำซ้ำเมื่อใหร่ ซึ่งเงื่อนไขที่จะหยุดการทำซ้ำของ ipPCA คือการทดสอบว่ากลุ่มประชากรย่อยที่ได้ทำการแบ่งออกมาจากกลุ่มประชากรทั้งหมดนั้น เป็นกลุ่มประชากรของบุคคลที่มีความคล้ายคลึงกันทางพันธุกรรม

ในแต่ละขั้นตอนการทำ PCA ผลที่ได้คือ eigenvalues และ eigenvectors ซึ่ง eigenvalues จะเป็นค่าที่ใช้สำหรับทดสอบว่ากลุ่มประชากรย่อยที่ได้เป็นกลุ่มที่มีความคล้ายคลึงกันทางพันธุกรรม โดยวิธีทดสอบนี้จะใช้การทดสอบค่าทางสถิติที่เรียกว่า Tracy-Widom (TW) [3] ส่วนค่าของ eigenvectors จะนำมาใช้สำหรับการคำนวนหาค่าสัมประสิทธ์เพื่อใช้ในการค้นหากลุ่มประชากรย่อยในแต่ละขั้นตอนการทำ PCA โดยการหากลุ่มประชากรย่อยในแต่ละขั้นตอนนั้นจะใช้ระเบียบวิธีการจัดกลุ่มโดยจะให้จะนวนของกลุ่มที่เป็นไปได้เท่ากับ 2 กลุ่มเสมอ โดยระเบียบวิธีสำหรับหารแยกกลุ่มนี้จะใช้แบบ fuzzy c-mean [4] ซึ่งเป็นระเบียบวิธีที่ไม่ซับซ้อนและให้ผลที่ดี

ขั้นตอนการแปลงข้อมูลจีโนไทป์

เนื่องจากข้อมูลจีโนไทป์เป็นข้อมูลที่เป็นเมตริกซ์ของอักขระดังแสดงในรูปที่ 1 ตามชนิดของอัลลีลของสนิปแต่ละตำแหน่ง ทางด้านแถวจะเป็นตัวอย่าง ส่วนทางด้านคอลัมน์จะเป็นสนิป จุดประสงค์ในการแปลงข้อมูลจีโนทัยป์เพื่อเอื้อต่อการคำนวณค่าในหัวข้อถัดไป เมตริกซ์อักขระจะถูกแปลงเป็น เมตริกซ์จำนวนเต็ม เพื่อให้ระเบียบวิธี PCA สามารถคำนวนได้ โดยค่าที่ใช้จะเป็นตัวเลขดังต่อไปนี้



รูปที่ 1 การแทนค่าข้อมูจีโนไทป์ด้วยตัวเลขจำนวนเต็ม

Homozygous Wild-type ซึ่งเป็นอัลลีลที่มีความถี่สูงที่สุดให้ค่า 0

Heterozygous ให้ค่าเป็น 1

Homozygous Mutant-type ให้ค่าเป็น 2

ข้อมูลที่ขาดหายไป (Missing-value Data) ให้ค่าเป็น -1

เราจะเรียกเมตริกซ์ข้อมูลที่ถูกแปลงค่านี้ว่าเมตริกซ์ X ดังสมการที่ 1

$$X = \overbrace{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{pmatrix}}^{\text{GenotypeData}} = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_N \end{pmatrix} \quad (1)$$

โดยที่ X คือเมตริกซ์ของข้อมูลสนิป

M คือจำนวนสนิป

N คือจำนวนตัวอย่าง

$\vec{x}$ คือเวคเตอร์ของสนิป

ข้อมูลที่ขาดหายไปอาจจะมีผลกับการวิเคราะห์หากลุ่มย่อยของประชากรได้ ซึ่งสามารถทดสอบได้โดยการแปลงข้องมูลจีโนไทป์ อื่น ๆ เป็น 0 และแปลงค่าของข้อมูลที่หายไปเป็น 1 จากนั้นให้ทำการวิเคราะห์ PCA และดูการกระจายตัวของสัมประสิทธ์ว่ามีการแยกกลุ่มออกมาหรือไม่ ถ้ามีกลุ่มแยกออกมาแสดงว่าข้อมูบที่หายไปมีผลกับการวิเคราะห์จำแนกประชากรกลุ่มย่อย ซึ่งอาจจะแก้ไขโดยการไม่รวมกลุ่มที่แยกออกมาไปในการวิเคราะห์

ขั้นตอนวิธีในการค้นหาโครงสร้างประชากรและจำนวนของประชากร (K)

ขั้นตอนนี้มีหน้าที่หาโครงสร้างของประชากรที่มีอยู่ในข้อมูลทั้งหมดรวมถึงจำนวนประชากร *K* โดยไม่ อาศัยข้อมูลป้ายกำกับ (labels) ซึ่งเป็นข้อมูลที่ได้มาตั้งแต่การเก็บตัวอย่างที่บ่งชี้ว่าตัวอย่างมาจากประชากรกลุ่มใด เพราะจำนวนของประชากรย่อยที่มีอยู่ในข้อมูลนั้นอาจมีมากกว่าหรือน้อยกว่าข้อมูลป้ายกำกับก็เป็นได้ ใน ขั้นตอนนี้เราจะอาศัยการวิเคราะห์พรินซิพัลคอมโพเนนต์หรือ Principal Component Analysis (PCA) เป็น เครื่องมือดังรูปที่ 3 ซึ่งอธิบายขั้นตอนในหัวข้อนี้ทั้งหมด

การวิเคราะห์พรินซิพัลคอมโพเนนต์หรือ Principal Component Analysis (PCA) เป็นระเบียบวิธีทาง คณิตศาสตร์ที่เกี่ยวข้องกับการใช้เทคนิคที่ชื่อว่าการกระจายเมตริกซ์แบบซิงกูลาร์ (Singular Value Decomposition) ทำการแปลงตัวแปรที่มีความสัมพันธ์กันหลายๆตัวในข้อมูล ไปยังตัวแปรที่ไม่สำพันธ์กันอีก จำนวนหนึ่งซึ่งมีจำนวนน้อยกว่าเดิมมาก เราเรียกตัวแปรนี้ว่า Principal component โดยวิธีการนี้เราจึงสามารถลด มิติของข้อมูลได้ ในขณะที่ความสำพันธ์ของตัวแปรเริ่มต้นยังคงถูกรักษาไว้ เราสามารถแสดงตัวแปรที่ถูกแปลง มายังโดเมนใหม่จากการวาดกราฟแบบ scatter plot แล้วสามารถแยกกลุ่มอย่างมีประสิทธิภาพด้วยขั้นตอนวิธีการ จัดกลุ่ม PCA สร้างโดเมนใหม่ได้จากการคำนวณเมตริกซ์โควาเรียนซ์ (Covariance Matrix) โดยใช้ข้อมูลสนิปที่ ถูกดัดแปลงให้เป็นตัวเลข 0, 1 และ 2 ตามสมการที่ 1

เนื่องจากเมตริกซ์โควาเรียนซ์โดยทั่วไปจะมีขนาดใหญ่ ทำให้อยากที่จะทำการคำนวณหาโดเมนใหม่ อย่างไรก็ตามเราสามารถใช้เทคนิคทางเมตริกซ์เข้ามาช่วยในการลดจำนวนครั้งของการคำนวณโดยการหาเมตริกซ์ ของ XX^T ซึ่งมีขนาดเล็กลงมากเมื่อเทียบกับการคำนวณเมตริกซ์โควาเรียนซ์ของ X โดยตรง อาศัยคุณสมบัติ ของ Singular value decomposition (SVD) ของเมตริกซ์ใดๆ ตามสมการดังต่อไปนี้

$$X = USV^T \tag{2}$$

จะได้ว่า

$$XX^T = US^2U^T \tag{3}$$

โดยที่ *U* คือ เมทริกซ์ยูนิแทรีด้านขวา

S คือ เมทริกซ์เฉียงที่มีจำนวนจริงที่ไม่เป็นลบอยู่ที่ตำแหน่งเฉียง

V คือ เมทริกซ์ยูนิแทรีด้านซ้าย

ดังนั้นเราสามารถคำนวนณหาคือ เมทริกซ์ยูนิแทรีด้านซ้าย *V* เพื่อสร้างโดเมนใหม่ได้จากสมการที่ 4

$$V = S(1:r, 1:r)^{-1/2}U(:, 1:r)^T X \tag{4}$$

โดยที่ r คือแรงค์ของเมตริกซ์ *S*

: คือสัญลักษณ์แทนการต่อเนื่องของดัชนีในแถวหรือหลักของเมตริกซ์

จากนั้นเราจะสามารถแปลงข้อมูลได้ดังสมการที่ 5

$$X_t = XV^T \tag{5}$$

โดยที่ X_t คือข้อมูลสนิปที่ถูกแปลงแล้ว โดยค่า X_t ที่ได้นี้ จะถูกนำมาหากลุ่มย่อยของประชากรโดยระเบียบวิธีการจัดกลุ่ม fuzzy c-mean ซึ่งจะทำการแบ่งกลุ่มของประชากรออกเป็น 2 กลุ่มย่อยและทำการทดสอบว่ากลุ่มประชากรย่อยที่ได้แต่ละกลุ่มมีความเป็นกลุ่มที่มาความคล้ายคลึงกันทางพันธุกรรมหรือไม่โดยใช้การทดสอบทางสถิติแบบ TW

การทดสอบทางสถิติแบบ TW จะใช้ค่าของ eigenvalue ซึ่งก็คือค่าในแนวเส้นทะแยงมุมของเมตริกซ์ $S^2 = diag[\lambda_1, \lambda_2, \ldots, \lambda_M]$ ซึ่งก็คือค่า eigenvalue ซึ่ง [3] ได้แสดงว่าค่าของ eigenvalue ที่ใหญ่ที่สุดเมื่อได้ทำการปรับค่าอย่างเหมาะสมแล้วจะมีค่าการกระจายตัวแบบ TW โดยที่ค่านี้จะถูกคำนวนดังขั้นตอนดังต่อไปนี้

$$\mu(M,N) = \frac{(\sqrt{N-1}+\sqrt{M})^2}{N} \tag{6}$$

$$\sigma(M,N) = \frac{(\sqrt{N-1}+\sqrt{M})}{N}\left(\frac{1}{\sqrt{N-1}}+\frac{1}{\sqrt{M}}\right)^{1/3} \tag{7}$$

ค่าทดสอบทางสถิติของ TW จะถูกคำนวนโดย

$$x = \frac{\lambda_1 - \mu(M,N)}{\sigma(M,N)} \tag{8}$$

อย่างไรก็ตามเพื่อที่จะให้ค่าทดสอบทางสถิติเป็น TW สมการที่ (8) จะถูกแทนที่โดย [3]

$$x = \frac{L_1 - \mu(M,N)}{\sigma(M,N)} \tag{9}$$

โดยที่

$$L_1 = \frac{M\lambda_1}{\sum_{i=1}^{M}\lambda_i} \tag{10}$$

ระเบียบวิธี ipPCA สามารถสรุปได้ดังต่อไปนี้

ขั้นตอนก่อนการวิเคราะห์

ให้ทำการทดสอบว่าข้อมูลที่หายไปมีผลต่อการวิเคราะห์ข้อมูลหรือไม่โดยทำการแทนค่าข้อมูลจีโนไทป์อื่น ๆ ด้วย 0 และแทนค่าข้อมูลที่หายไปด้วย 1 จากนั้นให้ทำการวิเคราะห์ PCA เพื่อทำการสังเกตุดูว่าค่าที่หายไปทำให้เกิดกลุ่มที่แยกออกมาจากข้อมูลอื่น ๆ หรือไม่ ถ้าเกิดมีกลุ่มแยกออกมา แสดงว่าข้อมูลที่หายไปจะมีผลกระทบกับการวิเคราะห์กลุ่มประชากรย่อย ซึ่งอาจจะแก้ไขได้โดยการไม่รวมข้อมูลที่แยกออกมาดังกล่าวเข้าไปในการวิเคราะห์ [3]

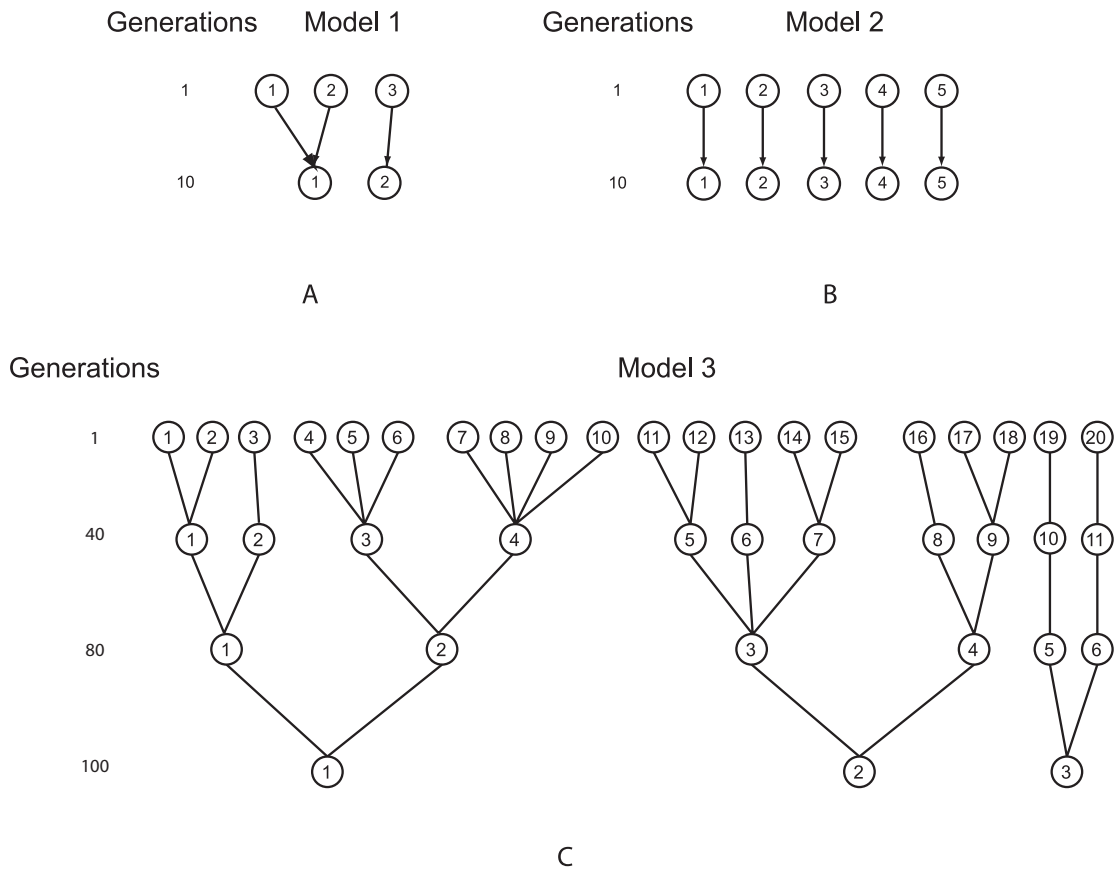
ขั้นตอนของ ipPCA

1. สร้างเมตริกซ์ X_i สำหรับข้อมูลที่ยังไม่ได้ทำการหยุด
2. ทำการใช้ SVD [5] เพื่อแยกหาค่า eigenvalue และ eigenvectors ตามสมการที่ (2) คือ $X_i X_i^T = U_i S_i V_i^T$
3. คำนวนค่าสัมประสิทธ์ของแต่ละบุคคลใน PCA เพื่อทำการแยกกลุ่มโดยใช้จำนวนของ principal components เท่ากันแรงค์ของเมตริกซ์
4. ทดสอบว่าหลังจากแยกกลุ่มแล้ว กลุ่มที่ได้มีความคล้ายกันทางพันธุกรรมหรือยัง โดยการใช้การทดสอบทางสถิติแบบ TW [3] ซึ่งจะทำการหยุดถ้าค่าทดสอบทางสถิติมีความสำคัญคือค่า p-value < 10^{-12} ถ้าค่า p-value สูงกว่านี้ก็ให้ทำการแยกกลุ่มโดยขั้นตอนต่อไป
5. ทำการแยกกลุ่มที่ไม่ได้ถูกหยุดออกเป็นสองกลุ่มโดยใช้ระเบียบวิธี fuzzy c-mean [6]
6. ทำซ้ำขั้นตอนที่หนึ่งจนกระทั่งทุก ๆ ข้อมูลมีความคล้ายคลึงกันทางพันธุกรรม
7. จำนวนของกลุ่มประชากรย่อย (*K*)สามารถหาได้โดยการนับจำนวนของกลุ่มที่มีความคล้ายคลึงกันทางพันธุกรรมทุก ๆ กลุ่ม

ขั้นตอนการทำ ipPCA นี้อาจจะต้องทำหลายครั้งเพื่อดูถูกต้องของผลการวิเคราะห์

บทที่ 3 การทดสอบกับข้อมูลจำลอง

เพื่อทำการทดสอบระเบียบวิธีที่ได้ทำการพัฒนาขึ้มใหม่ ข้อมูลของกลุ่มประชากรได้ถูกจำลองขึ้นมาโดยใช้ซอฟแวร์ GENOME [7] โดยซอฟแวร์ GENOME จะเป็นการสร้างประชากรจำลองแบบย้อนกลับเวลา ipPCA ได้ถูกทำการทดสอบด้วยข้อมูลประชากรจำลองจำนวน 3 ชุด โดยมีรายละเอียดดังนี้ ข้อมูลชุดแรกจะเป็นแบบที่มีการผสมกันของกลุ่มประชากรแต่มีจำนวนกลุ่มประชากรไม่มากคือ 3 กลุ่ม ข้อมูลชุดที่ 2 จะเป็นข้อมูลที่ไม่มีการผสมกันของกลุ่มประชากรโดยมีจำนวนกลุ่มประชากรย่อยทั้งหมด 5 กลุ่มข้อมูลชุดที่ 3 เป็นข้อมูลที่มีความสลับซับซ้อนสูงโดยมีจำนวนของกลุ่มประชากรย่อย ทั้งหมด 20 กลุ่มและมีการผสมกันระหว่างกลุ่มประชากร ข้อมูลทั้ง 3 ชุดนี้จะเรียกว่า Model 1, Model 2 และ Model 3 ตามลำดับ โดยโครงสร้างของประชากรทั้ง 3 กลุ่ม แสดงในรูปที่ 2



รูปที่ 2 โครงสร้างของกลุ่มประชากรที่ทำการทดสอบ

พารามิเตอร์ที่ใช้ในการรัน GONOME มีดังต่อไปนี้

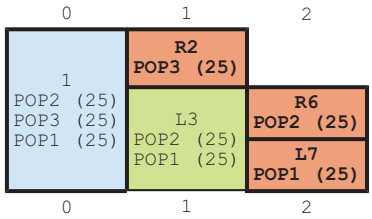
The model 1 dataset parameters:
-pop 3 50 50 50 –c 20 –s 500 –N model1.txt

The model 2 dataset parameters:
-pop 5 50 50 50 50 50 –c 20 –s 500 –N model2.txt

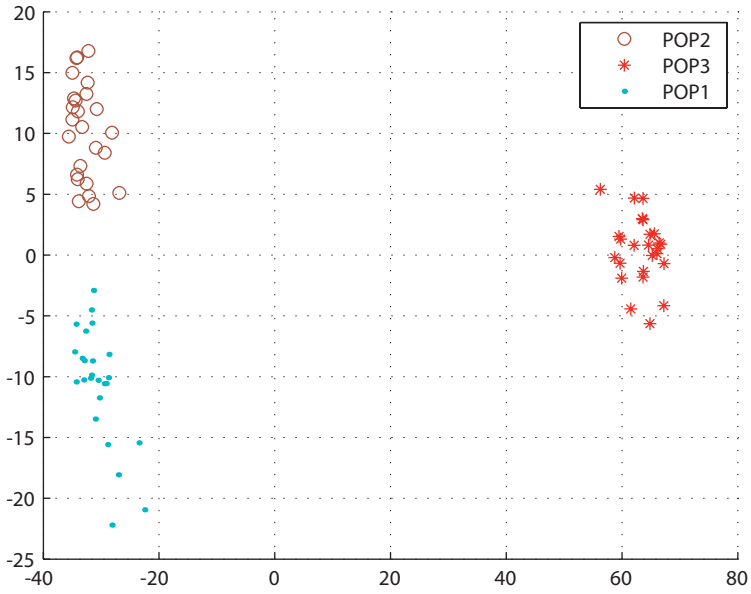
The model 3 dataset parameters:
-pop 20 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 -c 20 –s 500 –N model3.txt

โดยพารามิเตอร์ข้างบนจะให้ข้อมูลหนาด 10000 สนิปซ์สำหรับแต่ละบุคคล โดยที่ข้อมูล Model 3 ได้ทำการจำลองทั้งหมด 30 ชุดเพื่อทำการทดสอบผลลัพท์ที่สอดคล้องกันของระเบียบวิธี ipPCA

เพื่อที่จะทดสอบความถูกต้องของระเบียบวิธี ipPCA ข้อมูลจำลองที่มีจำนวนกลุ่มประชากรย่อยไม่มากคือ 3 และ 5 กลุ่มได้ถูกนำมาวิเคราะห์โดยระเบียบวิธี ipPCA ซึ่งสำหรับข้อมูลนี้ระเบียบวิธี ipPCA ทำการหาจำนวนของกลุ่มประชากรย่อยได้อย่างถูกต้องสอดคล้องกับค่าของพารามิเตอร์ที่ใช้ในการจำลอง นอกจากนี้ผลการหาว่าแต่ละบุคคลอยู่ในกลุ่มประชากรย่อยใดก็เป็นไปอย่างถูกต้อง ดังแสดงในรูปที่ 3 และ 4 ซึ่งเป็นผลของกลุ่มประชากรจากการทดสอบวิเคราะห์ด้วย ipPCA จำนวน 10 ครั้ง สำหรับกลุ่มประชากรที่ซับซ้อน (Model 3) ระเบียบวิธี ipPCA ได้ถูกทดสอบกัลข้อมูลจำลองของ Model 3 จำนวน 30 กลุ่มข้อมูล โดยที่ผลการทดสอบแสดงให้เห็นว่าระเบียบวิธี ipPCA สามารถกาจำนวนกลุ่มได้อย่างสอดคล้องกันในทุก ๆ กลุ่มข้อมูล นอกจากนี้ยังสามารถจำแนกบุคคลเข้าสู่กลุ่มประชากรย่อยได้อย่างสอดคล้องกันด้วย โดยที่จำนวนกลุ่มประชากรที่หาได้มีค่า mode เท่ากับ 20 โดยที่มีค่าอยู่ระหว่าง 19 ถึง 22 และมีอัตราการผิดพลาดของการจำแนกบุคคลเฉลี่ยอยู่ที่ 6.07% โดยทีอัตตราผิดพลาดมีค่าอยู่ระหว่าง 1.9% - 17.5%

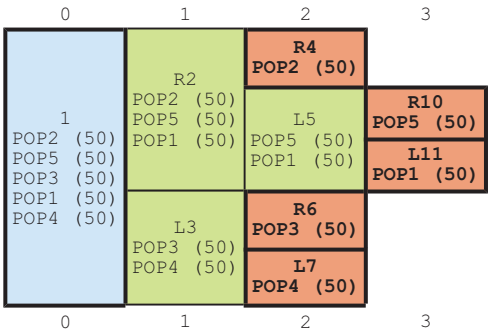


A

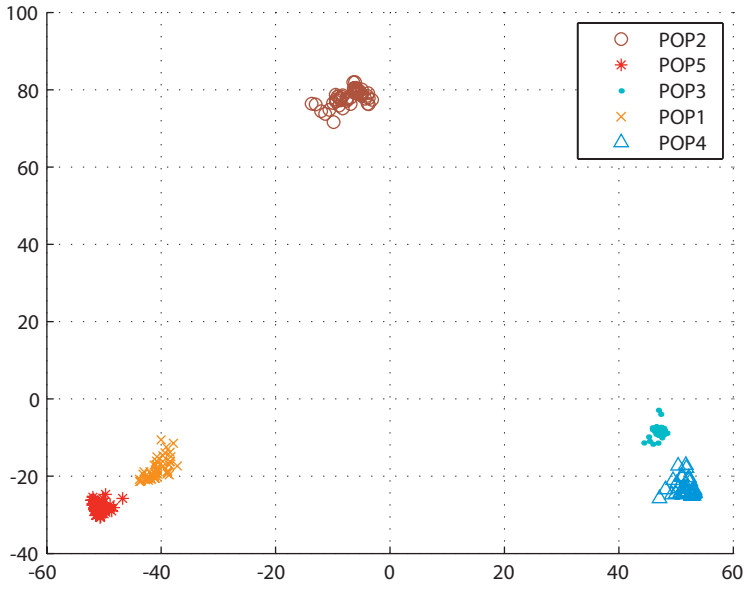


B

รูปที่ 3 ผลการทดสอบการวิเคราะห์ข้อมูลประชากรด้วย ipPCA จำนวน 10 ครั้งโดยผลที่นำมาแสดงเป็นผลจากผลการวิเคราะห์ส่วนใหญ่ A) แผนภูมิต้นไม้ของประชากรโดยแต่ละช่องคือกลุ่มประชากรย่อย กลุ่มประชากรย่อย SP-1, SP-2 และ SP-3 คือป้ายข้อมูลของประชาการย่อยที่ใช้ในการจำลอง B) เป็นผลการพลอตแบบกระจายของแต่ละบุคคลใน PC1 และ PC2 โดยแต่ละจุดแทนแต่ละบุคคลในกลุ่มประชากร



A



B

รูปที่ 4 ผลการทดสอบการวิเคราะห์ข้อมูลประชากรด้วย ipPCA จำนวน 10 ครั้งโดยผลที่นำมาแสดงเป็นผลจากผลการวิเคราะห์ส่วนใหญ่ A) แผนภูมิต้นไม้ของประชากรโดยแต่ละช่องคือกลุ่มประชากรย่อย กลุ่มประชากรย่อย SP-1, SP-2, SP-3, SP-4 และ SP-5 คือป้ายข้อมูลของประชาการย่อยที่ใช้ในการจำลอง B) เป็นผลการพลอตแบบกระจายของแต่ละบุคคลใน PC1 และ PC2 โดยแต่ละจุดแทนแต่ละบุคคลในกลุ่มประชากร

บทที่ 4 การทดสอบกับข้อมูลจริง

เพื่อที่จะทดสอบว่าระเบียบวิธีใหม่นี้มีประสิทธิภาพกับข้อมูลจริงเป็นอย่างไร จึงได้ทำการทดสอบกับข้อมูลจริงจำนวน 3 ชุดข้อมูล โดยทำการเปรียบเทียบกับระเบียบวิธี STRUCTURE [2] และ AWclust [8] โดยโปรแกรม STRUCTURE เป็น version 2.2 โดยใช้พารามิเตอร์ดังต่อไปนี้ 100000 burn-ins 100000 runs with admixture model, no LD model โปรแกรม AWclust ถูใช้วิเคราะห์ข้อมูลโดยใช้พารามิเตอร์ที่มากับโปรแกรม

ระเบียบวิธี ipPCA ได้ถูกทำการทดสอบกับข้อมูลจริงจำนวน 3 ข้อมูลคือ กล่าวคือ HapMap, Bovine (วัว) และข้อมูลของ Shriver โดยที่ข้อมูลของ HapMap ได้มากจาก [9] โดยที่ข้อมูล HapMap ประกอบไปด้วยกลุ่มประชากรย่อยจำนวน 4 กลุ่มคือ Han Chinese from Beijing (CHB), Japanese from Tokyo (JPT), Caucasian European from Utah (CEU), and Yoruba from Ibadan (YRI) โดยมีจำนวนสนิปซ์ทั้งหมด 1533661 สนิปซ์ ในการวิเคราะห์ข้อมูลของ HapMap จะแบ่งออกเป็น 2 แบบคือใช้จำนวนสนิปซ์ 50000 สนิปซ์โดยทำการเลือกมากจาก 1533661 สนิปซ์ และ ในส่วนที่สองคือใช้จำนวนสนิปซ์ทั้งหมด 1533661 สนิปซ์ โดยที่ก่อนทำการวิเคราะห์ได้มีการทำขบวนการก่อนการวิเคราะห์และได้ทำการเอาบุคคล 2 คนที่เป็น JPT ออกจากข้อมูล

สำหรับข้อมูลวัวจะ download จาก [10] ซึ่งจะประกอบไปด้วยจำนวนสนิปซ์ทั้งหมด 9329 สนิปซ์ หลังจากที่ได้ทำการวิเคราะห์แล้วสามารถตรวจสอบข้อมูลที่แตกต่างได้ จึงทำการวิเคราะห์ข้อมูลทั้งหมด สำหรับข้อมูลประชากรของ Shriver ประกอบไปด้วยจำนวนคนทั้งสิ้น 307 คน จาก 14 กลุ่มประชากรย่อยที่ต่างกันทางภาษาและภูมิภาค โดยข้อมูลทั้งหมดได้รับความอนุเคราะห์จาก Prof. Mark D. Shriver ซึ่งเป็นข้อมูลใหม่ล่าสุดจากผลงานที่ตีพิมพ์ [11] ข้อมูลนี้ประกอบไปด้วย 11555 สนิปซ์สำหรับแต่ละบุคคลซึ่งกระจายทั่วฃจีโนม หลังจากที่ได้ทำการทดสอบแล้ว ข้อมูลชุดนี้ไม่มีบุคคลที่ต่างจากพวก จึงทำการวิเคราะห์ข้อมูลทั้งหมด

หลังจากที่ข้อมูล HapMap ได้ถูกวิเคราะห์โดย ipPCA แต่ละบุคคลที่แตกต่างกันตามภูมิภาคได้ถูกแยกอยู่ในกลุ่มประชากรย่อย 4 กลุ่มโดยที่แต่ละกลุ่มจะมีแต่เฉพาะคนที่มีสัญลักษณ์เดียวกัน (รูปที่ 5) กลุ่มประชากรย่อย YRI สามารถตรวจเจอได้ในรอบแรกของการวิเคราะห์ ipPCA ขณะที่ กลุ่มประชากรย่อย CEU สามารถตรวจเจอได้ในรอบที่ 2 และสุดท้ายกลุ่มประชากรย่อย CHB และ JPT อยู่ในรอบที่ 3 สำหรับการวิเคราะห์ข้อมูลขนาด 1533661 สนิปซ์นั้นให้ผลเหมือนกับการวิเคราะห์ข้อมูลที่ 50000 สนิปซ์

สำหรับข้อมูล Bovine ประกอบไปด้วยประชากรวัวจาก 9 สายพันธุ์ที่มาจากถิ่นที่อยู่ที่ต่างกัน โดยที่ Brahman (BRM) สืบเชื้อสายมาจาก Zebu (หรือเรียกอีกอย่างหนึ่งว่า B. indicus) ซึ่งเป็นสายพันธุ์ที่ต่างจากสายพันธุ์ Taurine มาก สายพันธุ์ Santa Gertrudis (SGT) เป็นสายพันธุ์ผสมระหว่าง BRM กับ Shorthorn ส่วนสายพันธุ์อื่น ๆ Angus (ANG), Charolais

(CHL), Limousin (LMS), Hereford (HFD), Narwegian red (NRC), Jersey (JER) และ Holstein (HOL) เป็นกลุ่ม Taurine ที่มีจุดกำเนิดมาจากยุโรปสำหรับรายละเอียดของสายพันธุ์สามารถดูได้จาก[12]

รูปการพลอตแบบกระจายของกลุ่มวัวทั้งหมดที่แสดงวัวแต่ล่ะตัว จะเห็นได้ว่ากลุ่มประชากรย่อย HFD, JER, HOL, LMS, ANG, CHL, และ NRC ได้รวมกันเป็นกลุ่มใหญ่โดยวัวทั้งหมดนี้มาจากกลุ่ม Taurine ขณะที่กลุ่ม BRM และ SGT ได้ปรากฏแยกออกมาต่างหาก ดังแสดงในรูปที่ 6 ในรูปที่ แสดงแผมภูมิต้นไม้ของกลุ่มประชากรย่อยจากผลที่เหมือนกันจากการวิเคราะห์ 10 ครั้งด้วย ipPCA ซึ่งการวิเคราะห์จะต้องการ ipPCA ทั้งหมด 6 รอบเพื่อที่จะได้เห็นกลุ่มประชากรย่อยทั้งหมด ซึ่งหลังจากการวิเคราะห์รอบที่ 2 กลุ่มประชากรย่อยของ BRM และ SGT ก็แยกออกมา นอกจากนี้กลุ่มประชากรย่อยที่ 3 ประกอบไปด้วยวัวจาก JER ซึ่งเป็นกลุ่มประชากรย่อยที่แยกออกมาจากกลุ่มของ taurine หลังจาที่ได้ทำการวิเคราะห์ด้วย ipPCA จำนวน 6 รอบแล้วก็จะได้จำนวนกลุ่มประชากรย่อยออกมาเพิ่มอีก 6 กลุ่มประชากรย่อยซึ่งแต่ละกลุ่มประกอบไปด้วยวัวที่มาจากสายพันธุ์เดียวกันเป็นส่วนใหญ่

สำหรับข้อมูลของ Shriver จากการทำ PCA รอบแรก จะเห็นได้ว่ากลุ่มประชากรย่อยส่วนใหญ่ได้รวมกันเป็นกลุ่มใหญ่ตามลักษณะ ทางภูมิศาสตร์โดยบางกลุ่มก็มีการซ้อนทับกันของประชากรกลุ่มย่อยอย่างที่แสดงในรูปที่ 8 หลังจากที่ได้ทำการวิเคราะห์ ipPCA สองรอบก็ปรากฏเห็นกลุ่มประชากรย่อยของ African Americans และ Puerto Rican หลังจากที่ได้ทำการวิเคราะห์ ipPCA หลายรอบมากขึ้น ก็ปรากฏให้เห็นกลุ่มประชากรย่อยมากขึ้น โดยที่แต่ละกลุ่มประชากรย่อยประกอบไปด้วยบุคคลที่มาจากเผ่าพันธุ์เดียวกันหรืออยู่ใกล้ ๆ กัน หลังจากที่ได้ทำการวิเคราะห์ ipPCA ไปทั้ง 6 รอบ ก็ไม่ปรากฏ กลุ่มประชากรย่อยอีก (รูปที่ 9) และการหาจำนวนกลุ่มประชากรย่อยนี้มีความสอดคล้องกันจากการวิเคราะห์ 10 ครั้ง มีเพียงครั้งเดียวที่ผลแตกต่างจากครั้งอื่น

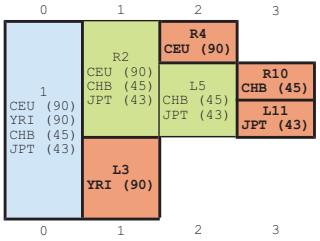
สำหรับการหาจำนวนกลุ่มประชากรย่อยและการนำแต่ละบุคคลเข้าไปอยู่ในกลุ่มประชากรย่อยที่เหมาะสม ที่ได้จาก ipPCA นั้นได้ทำการเปรียบเทียบกับระเบียบวิธีอื่น ๆ ที่มีอยู่ โดยผลการเปรียบเทียบแสดงในตารางที่ 1 ซึ่งแสดงการเปรียบเทียบผลของการหากลุ่มประชากรย่อยจากข้อมูลทั้งหมด 35 ข้อมูล ระเบียบวิธีทั้งหมดให้ผลตรงกับเฉพาะข้อมูลที่ไม่ซับซ้อน (Model 1) สำหรับข้อมูลอื่น ๆ แต่ละระเบียบวิธีก็ให้ผลต่างกันไป เนื่องจากแต่ละระเบียบวิธีให้ผลของจำนวนกลุ่มประชากรไมตรงกัน จึงได้ทำการตรวจสอบการนำบุคคลเข้าสู่กลุ่มประชากรย่อยที่ได้

สำหรับข้อมูลที่ไม่ซับซ้อนการนำบุคคลเข้าสู่กลุ่มประชากรย่อยมีความสอดคล้องกันของระเบียบวิธี STRUCTURE, AWclust และ ipPCA สำหรับข้อมูลที่มีความซับซ้อนและมีความไม่สอดคล้องกันของจำนวนกลุ่มประชากรย่อย ผลระหว่างระเบียบวิธี STRUCTURE และ ipPCA ได้ถูกเปรียบเทียบ สำหรับข้อมูลที่มีความ ซับซ้อนมากคือข้อมูลของ Shriver โดยการเปรียบเทียบได้ใช้จำนวนกลุ่มประชากรย่อยที่ดีที่สุดของแต่ละระเบียบวิธี (รูปที่ 10) แต่ละกลุ่ม

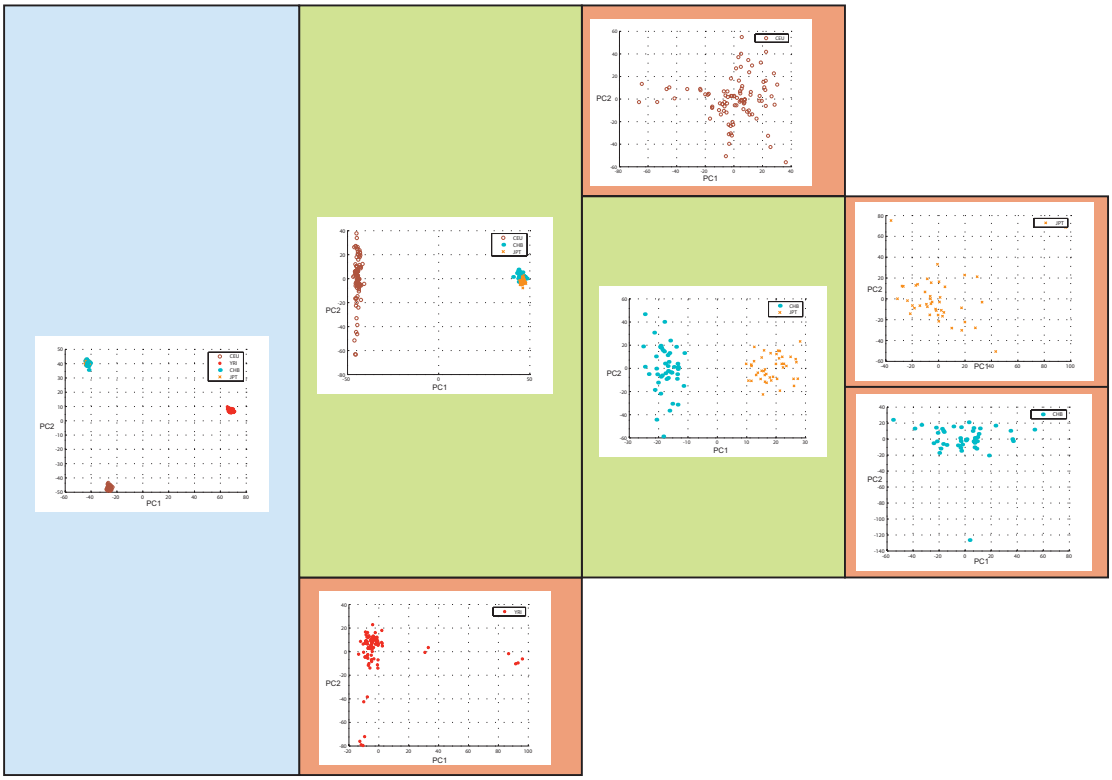
ประชากรย่อยจำนวน 12 กลุ่มที่หาได้จากระเบียบวิธี ipPCA จะมีจำนวนบุคคลในแต่ละกลุ่มใกล้เคียงกันคือจาก 19 ถึง 46 คน นอกจากนี้ กลุ่มบุคคลที่อยู่ในกลุ่มประชากรย่อย SP-4, SP-5, SP-6, SP-7, SP-8, SP-9, SP-11 และ SP-12 เป็นบุคคลที่มาจากสัญลักษณ์ของประชากรเดียวกัน สำหรับ 4 กลุ่มที่เหลือคือ SP-1, SP-2, SP-3 และ SP-10 ประกอบด้วยบุคคลที่มาจากกลุ่มประชากรมากกว่า 1 สัญลักษณ์ขึ้นไป

สำหรับกลุ่มประชากรย่อยที่ได้จากระเบียบวิธี STRUCTURE นั้นมีความแตกต่างจาก ipPCA ค่อนข้างมาก คือมีจำนวนกลุ่มประชากรย่อยทั้งหมด 14 กลุ่ม และจำนวนบุคคลที่อยู่ในแต่ละกลุ่ม โดยสองกลุ่มประชากรย่อยไม่มีบุคคลใดอยู่ ห้ากลุ่มประชากรย่อยมีบุคคลอยู่น้อยกว่า 10 คน และมีสองกลุ่มประชากรที่มีบุคคลอยู่มากกว่า 50 คน สังเกตุได้ว่าระเบียบวิธี STUCTURE ได้รวมกลุ่มออกเป็นสามกลุ่มแยกออกตามแหล่งที่อยู่ต่างกันคือ African รวมกับ African American และ Puerto Rican โดยที่ ipPCA ได้แยกสามกลุ่มนี้ออกจากกันในกลุ่ม SP-4, SP-8 และ SP-9 ตามลำดับ นอกจากนี้ STRUCTURE ได้แยกกลุ่มของ Caucasian, Puerto Rican, South Altaian และ West African ออกเป็นกลุ่มย่อยหลาย ๆ กลุ่มประชากรย่อยมากกว่า ipPCA และ STRUCTURE ยังได้ทำการรวมบุคคลจาก Nahua และ Quechua/Peru เป็นกลุ่มประชากรย่อยเดียวกัน และบุคคลจาก Asian กับ Aouth Altaian รวมเป็นกลุ่มประชากรย่อยเดียวกัน โดยที่ ipPCA สามารถแยกกลุ่มประชากรย่อยเหล่านี้ได้สอกคล้องกับสัญลักษณ์ของบุคคล (SP-5, SP-6, SP-11 และ SP-12)

เพื่อที่จะดูความแตกต่างระหว่างการจัดกลุ่มแบบไม่ทำซ้ำกับที่ทำซ้ำ จึงทำการเปรียบเทียบระหว่างสองวิธีนี้ โดยวิธีที่ไม่ได้ทำซ้ำจะเรียกว่า non-iterative (แถวที่ 3 ในรูปที่ 10) ผลของการจัดกลุ่มด้วยวีธี fuzzy c-mean โดยมีจำนวนของกลุ่มประชากรย่อยทั้งหมด 12 กลุ่มซึ่งเท่ากับจำนวนของกลุ่มประชากรย่อยที่ได้จาก ipPCA โดยที่จำนวนกลุ่มนี้ไม่ได้ถูกหาด้วยวิธีอื่นเนื่องจากเป็นข้อมูลที่มีความซับซ้อน Gap Statistic ไม่สามารถใช้หาจำนวนกลุ่มได้ ผลที่ได้แสดงให้เห็นว่าการไม่ทำซ้ำทำให้ผลแตกต่างกันอย่างมากของการนำบุคคลเข้าสู่กลุ่มประชากรย่อย ตัวอย่างเช่น มีสามกลุ่มที่มีคนอยู่น้อยกว่า 10 คน (3 Puerto Ricans, 8 East Africans, และ 4 New Guinea) และมีกลุ่มที่รวมกันเป็นกลุ่มใหญ่กลุ่มเดียวโดยมีคนที่อยู่ในกลุ่มนั้นมากถึง 68 คน (4 Spanish, 11 South Indian Mala, 11 South Indian Brahman และ 42 Caucasians) นอกจากนี้ บุคคลจาก Asian และ South Altaian ได้ถูกนำไปรวมกันเป็นกลุ่มเดียว ขณะที่คน East African และ New Guinea ได้แยกออกเป็นสองกลุ่มประชากรย่อย



A

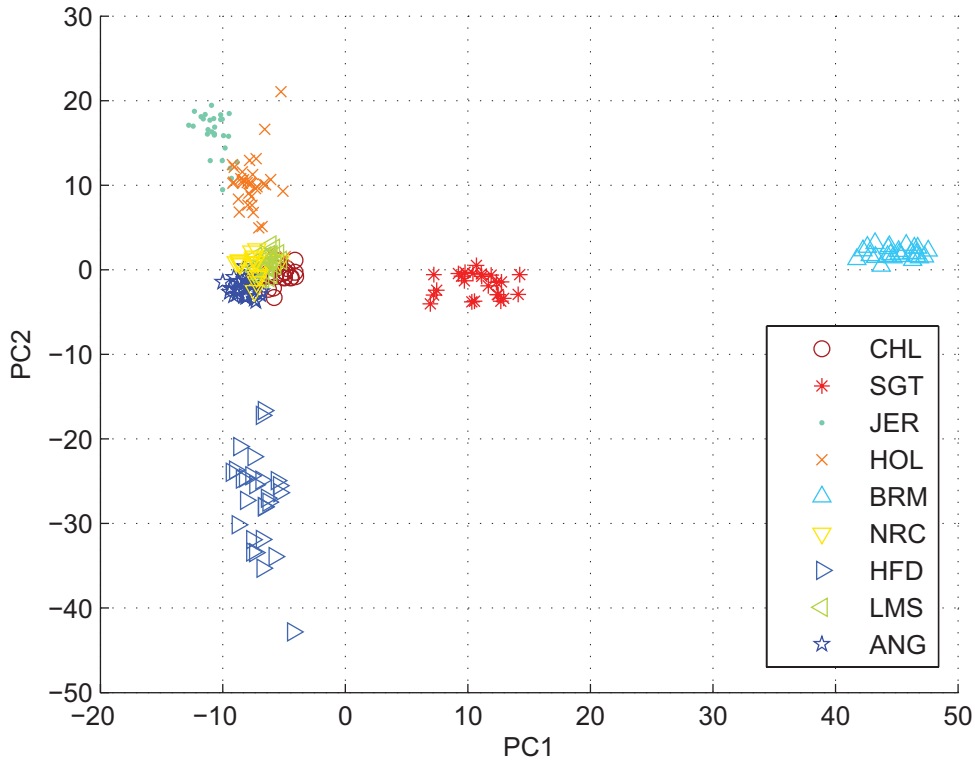


B

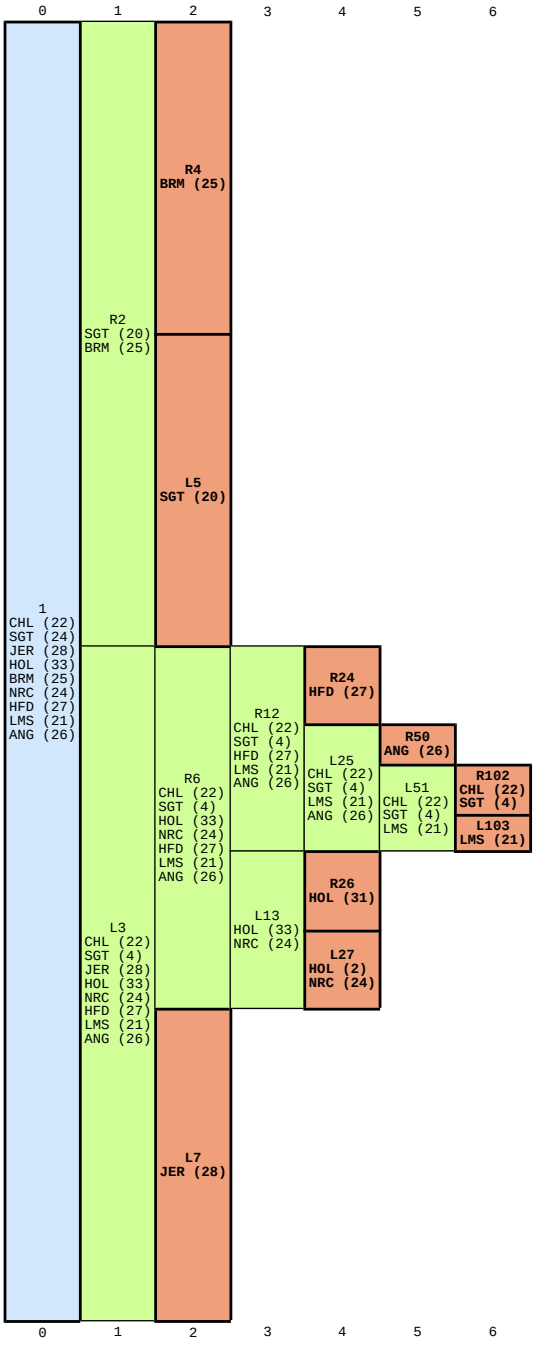
รูปที่ 5 ผลการวิเคราะห์ ipPCA ของข้อมูล HapMap

A) แผนผังต้นไม้ของกลุ่มประชากรย่อย โดยที่แต่ละช่องจะมีสัญลักษณ์ของกลุ่มประชากรย่อย จำนวนของบุคคลแสดงเป็นตัวเลขอยู่ในวงเล็บ และจำนวนของ PC ที่ใช้สำหรับการจักกลุ่มแสดงอยู่ในแต่ละช่อง ช่องสีน้ำเงินแสดงข้อมูลเริ่มต้น ส่วนข้อมูลที่ยังไม่ได้อยุดวิเคราะห์แสดงในช่องสีเขียว ขณะที่ช่องสีแดงคือกลุ่มประชากรย่อยที่หาได้โดย ipPCA

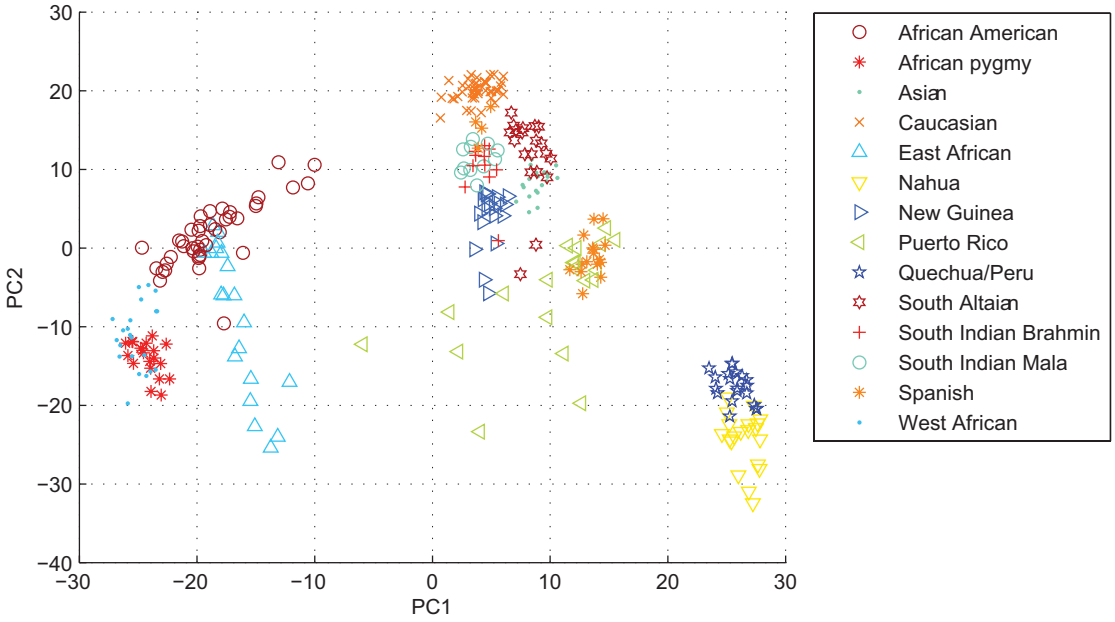
B) การพลอตแบบกระจายโดยใช้ PC1 และ PC2 โดยที่แต่ละจุดคือแต่ละบุคคล สัญลักษณ์ของแต่ละบุคคลแสดงอยู่ในช่อง



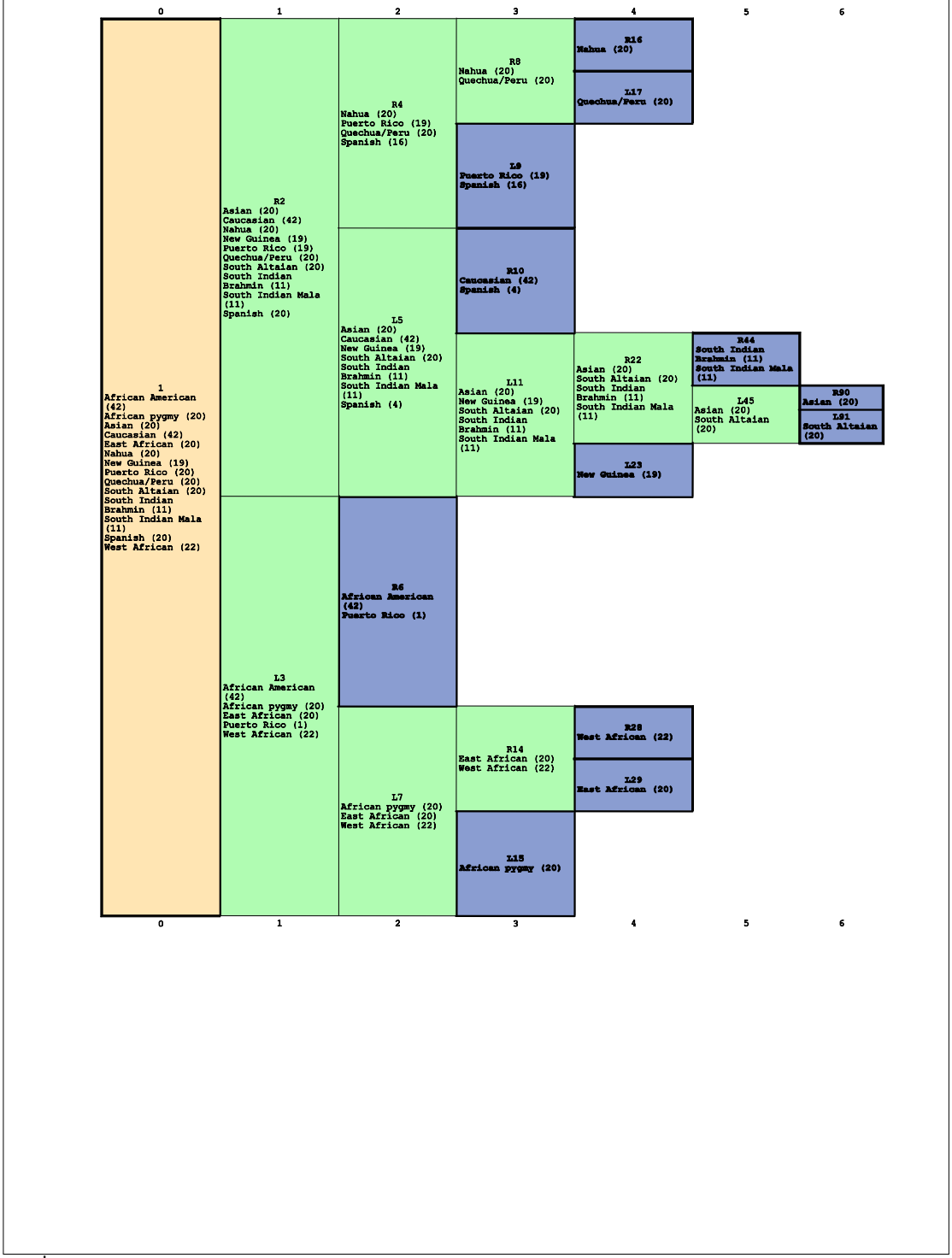
รูปที่ 6 การพลอตแบบกระจายตัวของข้อมูล Bovine ที่การวนรอบที่ศูนย์ โดยการพลอตนี้ได้ใช้ PC1 และ PC2 แต่ละจุดคือวัวแต่ละตัวโดยที่สายพันธุ์ของวัวสามารถดูได้จากรูป



รูปที่ 7 แผนภูมิต้นไม้ของประชากรของข้อมูล Bovine โดยแต่ละช่องจะแสดงสาพันธุ์วัวด้วยตัวย่อ CHL, SGT, JER, HOL, BRM, NRC, HFD, LMS, หรือ ANG จำนวนของวัวในแต่ละกลุ่มประชากรย่อยแสดงเป็นตัวเลขต่อจากสัญลักษณ์ ช่องสีแดงคือข้อมูลที่ยังไม่ได้ทำกาวิเคราะห์ สีเขียวคือ ข้อมูลที่ไม่ได้หยุด ส่วนช่องสีแดงคือช่องที่หยุดแล้ว



รูปที่ 8 การพลอตแบบกระจายตัวของข้อมูลของ Shriver ที่การวนรอบที่ศูนย์ โดยการพลอตนี้ได้ใช้ PC1 และ PC2 แต่ละจุดคือวัวแต่ละตัวโดยที่ชาติพันธุ์ของคนสามารถดูได้จากรูป



รูปที่ 9 แผนภูมิต้นไม้ของประชากรของข้อมูลของ Shriver โดยแต่ละช่องจะแสดงชาติพันธุ์ African American, African Pygmy, Asian, Caucasian, East African, Nahua, New Guinea, Puerto Rican, Quechua/Peru, South Altaian, South Indian Brahmin, South Indian Mala, Spanish หรือ West African จำนวนของคนในแต่ละกลุ่มประชากรย่อยแสดงเป็นตัวเลขต่อจากสัญลักษณ์ ช่องสีแดงคือข้อมูลที่ยังไม่ได้ทำกาวิเคราะห์ สีเขียวคือ ข้อมูลที่ไม่ได้หยุด ส่วนช่องสีแดงคือช่องที่หยุดแล้ว

ตารางที่ 1 เปรียบเทียบผลการจัดกลุ่มประชากรของระเบียบวิธีต่าง ๆ (AWclust, STRUCTURE, และ ipPCA) สำหรับการอนุมานหาจำนวนกลุ่มประชากรย่อย

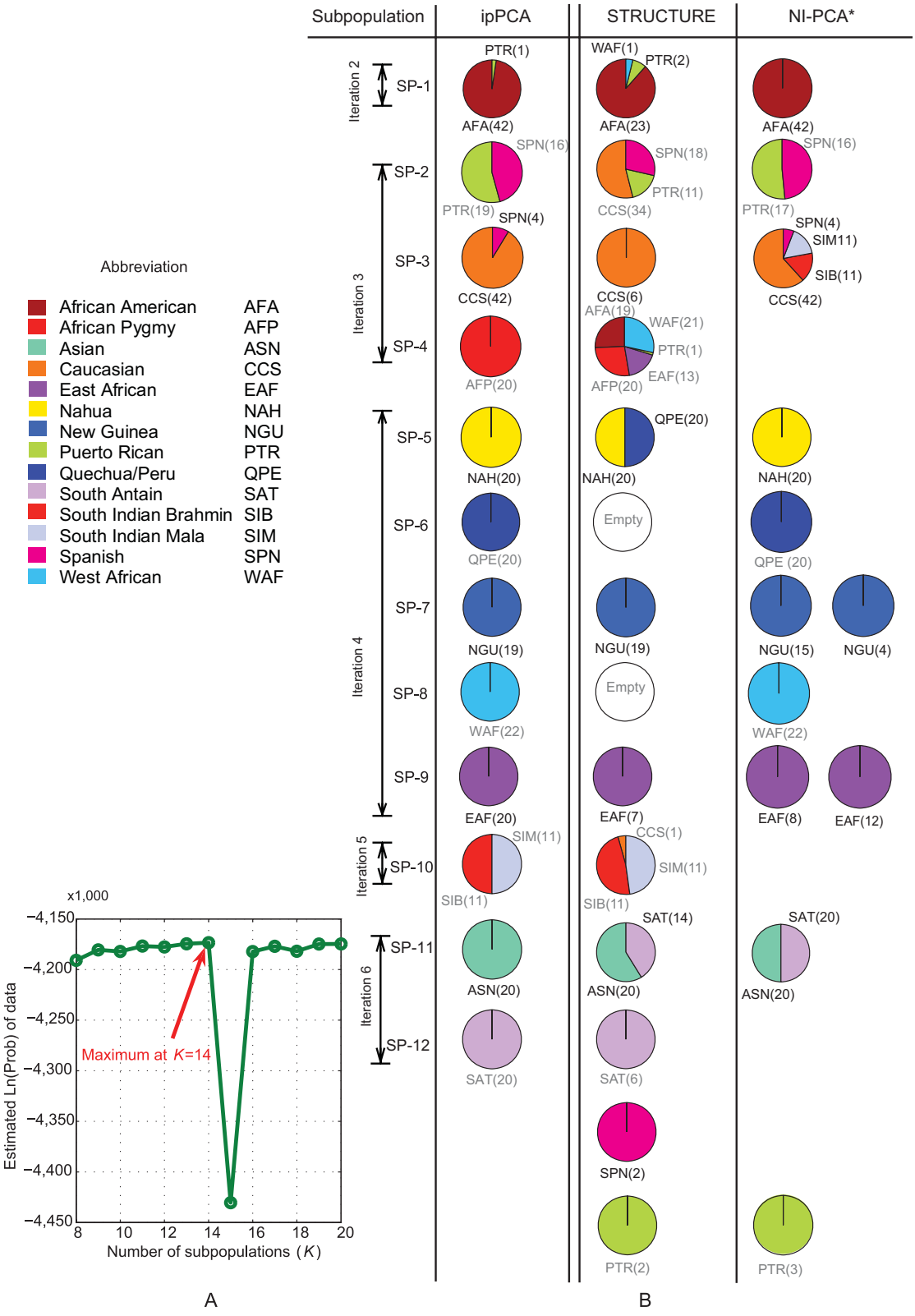
Dataset	AWclust	STRUCTURE	ipPCA
model 1 simulated (3 subpopulations)	3	3	3
model 2 simulated (5 subpopulations)	5	6	5
model 3 simulated (20 subpopulations, 30 datasets)	N/A[1]	N/D[2]	20[3]
HapMap (4 population labels with 50000 SNPs)	3	3	4
bovine (9 population labels with 9329 SNPs)	N/A[1]	10	9
Shriver's (14 population labels with 11555 SNPs)	N/A[1]	14[4]	12

[1] Not applicable since AWclust gap statistics limits $K < 8$

[2] Not done owing to the STRUCTURE computational constraint

[3] Modal value from 30 simulated datasets

[4] There are two subpopulation groups with no individuals assigned to them



รูปที่ 10 เปรียบเทียบการระบุบุคคลเข้าสู่กลุ่มประชากรย่อยของระเบียบวิธี ipPCA, STRUCTURE และ non-iterative PCA

A) Log probability ของการอนุมานหาจำนวนกลุ่มประชากรย่อย

B) เปรียบเทียบการระบุบุคคลเข้าสู่กลุ่มประชากรย่อย

บทที่ 5 วิเคราะห์และสรุปผล

การเปรียบเทียบระเบียบวิธี ipPCA กับระเบียบวิธีอื่น ๆ

ในงานวิจัยนี้เราได้แสดงให้เห็นว่าระเบียบวิธี ipPCA สามารถจำแนกประชากรได้อย่างถูกต้องสำหรับข้อมูลประชากรที่ซับซ้อนและไม่ซับซ้อน โดยระเบียบวิธีนี้สามรถหาจำนวนกลุ่มประชากรย่อยได้อย่างถูกต้องและยังสามารถหาได้ว่าบุคคลในกลุ่มประชากรอยู่ในกลุ่มประชากรย่อยใด นอกจากนี้ระเบียบวิธี ipPCA ยังสามารถคำนวนได้อย่างรวดเร็วและยังสามารถนำไปใช้ได้กับข้อมูลขนาดใหญ่ซึ่งไม่เหมือนระเบียบวิธีอื่นที่ไม่สามมาถใช้ได้กับข้อมูลที่มีขนาดใหญ่มากได้ ตัวอย่างเช่นระเบียบวิธี AWclust ไม่สามารถใช้กับข้อมูลที่มีจะนวนกลุ่มประชากรย่อยได้มากกว่า 7 และระเบียบวิธี STRUCTURE จะใช้เวลานานมากสำหรับข้อมูลที่มีจำนวนสนิปซ์มาก สำหรับข้อมูลชุดที่สามารถถูกวิเคราะห์ได้ด้วยระเบียบวิธี AWclust และ STRUCTURE นั้น ผลของการวิเคราะห์หาจำนวนกลุ่มประชากรย่อยด้วยระเบียบวิธีดังกล่าวสอดคล้องกันกับ ipPCA สำหรับข้อมูลที่มีจำนวนสนิปซ์มากและมีความซับซ้อนมากระเบียบวิธี ipPCA สามารถวิเคราะห์ข้อมูลได้เร็วกว่าระเบียบวิธี AWclust และ STRUCTURE มาก ตัวอย่างเช่นข้อมูลประชากรของ Shriver นั้น STRUCTURE ใช้เวลาในการวิเคราะห์ประมาณ 3 วันในขณะที่ ipPCA ใช้เวลาน้อยกว่า 10 นาทีในการวิเคราะห์ข้อมูลชุดเดียวกัน บนเครื่อง 32-core AMD 2.3 GHz Opteron โดยมีหน่วยความจำ 64 GByte และใช้ CentOS เป็นระบบประฏิบัติการ

การปรับปรุงการจัดกลุ่มด้วยการเลือก Principal Component (PC) ที่เหมาะสม

จากการสังเกตุจะเห็นว่าในการทำซ้ำรอบแรก ๆ จะต้องการจำนวน PC เป็นจำนวนมากกว่ารอบหลัง ๆ ในการแบ่งกลุ่ม (รูปที่ 3, 4, 5, 7 และ9) ซึ่งผลที่ได้นี้สอดคล้องกับผู้แต่ง EIGENSTRAT/SmartPCA ซึ่งแสดงว่าจำนวนของ eigenvector ที่สำคัญมีความสัมพันธุ์กับจำนวนของกลุ่มประชากรย่อย [3, 13] อย่างไรก็ตามจำนวนของกลุ่มประชากรย่อยไม่สามารถหาได้ง่าย ๆ จากจำนวนของ PC ที่สำคัญ เพราะว่าจำนวนของ PC ที่สำคัญเปลี่ยนไปในแต่ละข้อมูล ในระเบียบวิธี ipPCA นั้นขั้นตอนการแบ่งกลุ่มนั้นจะใช้จะนวนของ PC เท่ากับ Rank ของเมตริกซ์ซึ่งจะเปลียนไปตามข้อมูลและขึ้นอยู่กับจำนวนของบุคคลในแต่ละรอบของการวิเคราะห์ หลังจากแต่ละรอบของการทำ ipPCA ข้อมูลจะมีความซับซ้อนน้อยลงดังนั้นนจำนวน PC ที่ใช้ในการแบ่งกลุ่มก็จะมีจำนวนน้อนลงไปด้วย

การอนุมานหาจำนวนกลุ่มประชากรย่อย

การหากลุ่มประชากรย่อยหรือ K ที่เหมาะสมกับกลุ่มประชากรจะขึ้นอยู่กับความถูกต้องของการหาว่าบุคคลอยู่ในกลุ่มประชากรย่อยใด ถ้าค่าความถูกต้องนี้มีค่าสูงการอนุมานหาจำนวนกลุ่มประชากรย่อยก็จะมีความถูกต้องด้วย ระเบียบวิธี STRUCTURE และ AWclust

มีค่าความถูกต้องนี้ต่ำกว่า ipPCA ดังนั้นการอนุมานหาจำนวนกลุ่มประชากรย่อยจึงไม่ค่อยถูกต้อง ตัวอย่างเช่น ทั้งระเบียบวิธี STRUCTURE และ AWclust อนุมานจำนวนกลุ่มประชากรย่อยของข้อมูล HapMap ได้เท่ากับ 3 กลุ่ม โดยที่ไม่สามารถแยกกลุ่มของ CHB และ JPT ออกจากกันได้ ซึ่งการที่ระเบียบวิธี ipPCA สามารถแยกทั้ง CHB และ JPT ออกจากกันได้นี้ก็สอดคล้องกลับผลที่รายงานใน [14] ซึ่งได้สามารถหาสนิปซ์ที่เหมาะสมที่จะทำการแยกประชากรทั้งสองกลุ่มนี้ออกจากกันได้

ระเบียบวิธี ipPCA สามารถหากลุ่มประชากรที่มีความคล้ายคลึงกันทางพันธุกรรมได้โดยดูจากการจะจายตัวของ eigenvalue ด้วยวิธีนี้กลุ่มประชากรย่อยจะถูกกำหนดโดยวิธีมาตราฐานซึ่งสามารถแสดงให้เห็นถึงกลุ่มประชากรที่มีความคล้ายคลึงกันทางพันธุกรรมโดยไม่คำนึงถึงข้อมูลอื่น ๆ อย่างไรก็ตามการกำหนดกลุ่มประชากรย่อยในปัจจุบันนี้ก็ยังไม่สมบูรณ์ [37] จากเหตุผลอันนี้ทำให้การหาจำนวนกลุ่มของประชากรย่อย เป็นเรื่องที่สำคัญและท้าทาย [15, 16]

ความถูกต้องของการกำหนดบุคคลไปยังกลุ่มประชากรย่อย

ในแง่ของการกำหนดบุคคลจะเห็นได้จากข้อมูลของ Shriver ได้ว่ามีความแตกต่างกันระหว่างระเบียบวิธี ipPCA และ STRUCTURE โดยที่ STRUCTURE ได้อนุมานจะนวนของกลุ่มประชากรย่อยได้ 14 กลุ่ม โดยที่การกำนดบุคคลโดย STUCTURE นั้นไม่สอดคล้องกับวิธีอื่น ๆ [11, 17] นอกจากนี้การกำหนดบุคคลยังไม่ค่อยจะมีความหมาย เช่น มีสองกลุ่มที่ STRUCTURE ไม่ได้กำหนดบุคคลไป และมีถึง 5 กลุ่มที่ STRCUTURE กำหนดบุคคลไปน้อยกว่า 10 คน (รูปที 9) โดยการใช้จำนวนกลุ่มประชากรย่อยเท่ากับ 12 กลุ่มอย่างที่อนุมานได้โดย ipPCA การกำหนดบุคคลเข้าสู่กลุ่มโดย STRUCTURE จะเปลี่ยนไปโดยที่จะไม่มีกลุ่มที่ว่างเหลืออยู่นอกจากนี้กลุ่ม south Altaian ก็ได้แยกออกมาต่างหาก อย่างไรก็ตามก็ยังมีความไม่สอดคล้องกันกับผลที่ได้ตัวอย่างเช่นกลุ่มประชากรของ pan-African

การจัดกลุ่มของ ipPCA นี้แตกต่างจากการจักกลุ่มโดยการใช้ PCA อื่น ๆ ตัวอย่างเช่น [18] ซึ่งทำการจัดกลุ่มโดยใช้ข้อมูลทั้งหมดและใช้จำนวนของ PC เป็นจำนวนมาก ถ้าข้อมูลของประชากรมีความซับซ้อนมากการทำเช่นนี้ไม่สามารถกำหนดบุคคลเข้าสู่กลุ่มประชากรย่อยได้อย่างถูกต้องไม่ว่าจะใช้ระเบียบวิธีการจัดกลุ่มใดก็ตาม (เช่น k-mean, soft k-mean, spectral k-mean) เพราะว่ากลุ่มประชากรที่มีความใกล้ชิดกันก็จะมาอยู่รวมกันในเสปซใหม่ที่สร้างขึ้น นอกจากนี้ปัญหาใหญ่ที่สุดของ NI-PCA คือการหาจำนวนกลุ่มประชากรย่อยที่เหมาะสมก่อนที่จะทำการจัดกลุ่ม ซึ่งงานวิจัยนี้ก็แสดงให้เห็นว่า NI-PCA ไม่สามรถกำหนดบุคคลไปยังกลุ่มประชากรย่อยได้อย่างถูกต้อง ถึงแม้ว่าจะใช้จำนวนของกลุ่มประชากรที่ถูกต้องที่ได้จาก ipPCA ก็ตาม สำหรับ ipPCA นั้นได้เลือกใช้ระเบียบวิธี fuzzy c-mean สำหรับการจัดกลุ่มเพราะว่าจำนวนกลุ่มที่จะทำการแยกในแต่ละรอบจะเท่ากับ 2 กลุ่มเสมอ เพราะฉะนั้นระเบียบวิธีการจัดกลุ่มที่ไม่ซับซ้อนก็มีประสอทธิภาพเพียงพอที่จะแบ่งกลุ่มได้ นอกจากนี้จุดศูนย์กลางของกลุ่มที่

ได้จาก fuzzy c-mean ก็มีความสอดคล้องกันเมื่อเทียบกับระเบียบวิธี k-mean ที่ใช้กันโดยทั่วไป

การหากลุ่มประชากรย่อยตามลักษณะของระยะทางทางพันธุกรรมและประวัติของกลุ่มประชากร

การวิเคราะห์แบบวนซ้ำโดยการตัดกลุ่มประชากรย่อยออกนี้ได้ทำการกำหนดกลุ่มของกลุ่มประชากรย่อยโดยสะท้อนถึงการวิวัฒนาการและจุดกำเนิดของประชากร ซึ่งสามารถเห็นได้จากการวิเคราะห์ข้อมูลจำลอง (รูปที่ 3) และจากข้อมูลจริง (รูปที่ 5 - 9) จะเห็นได้ว่ากลุ่มประชากรที่ใกล้กันจะแยกออกหลังจากกลุ่มประชากรที่ใกลกว่าได้ถูกแยกออกไปแล้ว การวิเคราะห์แบบวนซ้ำโดยการตัดกลุ่มประชากรย่อยออกนี้ได้ช่วยให้การแยกกลุ่มประชากรสามารถทำได้อย่างมีระบบสอดคล้องกับความเกี่ยวข้องกันของกลุ่มประชากร กลุ่มประชากรสามารถแยกได้โดยการแยกเอากลุ่มประชากรที่อยู่ใกลกว่าออกไปก่อนจากนั้นจึงแยกกลุ่มประชากรที่เหลืออยู่ออกจากกัน ดังนั้นวิธีนี้จึงสามารถหากลุ่มประชากรย่อยที่ไม่สามารถหาได้โดยวิธีอื่นได้

สำหรับข้อมูลที่มีการผสมกันระหว่างกลุ่มประชากรย่อยนั้น ระเบียบวิธี ipPCA ก็สามารถแยกประชากรที่ผสมกันนี้ออกจากประชากรย่อยกลุ่มอื่นได้ เช่น กลุ่มประชากรย่อยสองกลุ่มที่ได้จากข้อมูล Bovine นั้นมีวัวที่มาจากฉลากที่ต่างกัน การที่วัว NRC และ HOL ทับกันนั้นก็ไม่ได้น่าแปลกใจเพราะว่ากลุ่ม NRC ไม่ได้เป็นสายพันธุ์ที่บริสุทธ์โดยสืบสายพันธุ์มาจาก HOL และ การที่มีวัวบางส่วนจากสายพันธุ์ SGT ได้ถูกกำหนดให้อยู่กับกลุ่มประชากรย่อยของกลุ่มสายพันธุ์หลัก taurine ก็แสดงให้เห็นว่าวัวสายพันธุ์ SGT นั้นมาจากสองสายพันธุ์คือ zebu กับ taurine

สำหรับข้อมูลของ Shriver ซึ่งมีความซับซ้อนมากกว่าของ Bovine นั้น ระเบียบวิธี ipPCA สามารถหาจำนวนกลุ่มประชากรย่อยได้ทั้งหมด 12 กลุ่ม และ แต่ละบุคคลได้ถูกกำหนดไปอยู่ในกลุ่มประชากรย่อยโดยสอดคล้องกับป้ายชื่อและ neighbor-joining tree ที่อยู่ใน [11] โดยมี 4 กลุ่มประชากรย่อยที่เป็นกลุ่มผสมโดยมีป้ายชื่อที่ต่างกันอยู่ในกลุ่มเหล่านี้ ในระหว่าง 4 กลุ่มนี้คน Puerto Rican ได้ถูกกำหนดไปอยู่ในกลุ่มประชากรย่อย 2 กลุ่ม การกำหนดบุคคลของกลุ่ม Puerto Rican นี้แสดงให้เห็นว่าจีโนมของ Puerto Rican ยังคงมีบรรพบุรุษเป็น European และ African [11, 19]

สรุป

ในงานวิจัยนี้ได้ทำการพัฒนาระเบียบวิธีใหม่โดยการปรับปรุงเพิ่มเติมขั้นตอนในการทำ PCA เพื่อให้สามารถทำการกากลุ่มประชากรได้อย่างถูกต้อง โดยวิธีการใหม่นี้สามารถระบุบุคคลไปยังกลุ่มประชากรย่อยได้อย่างวเหมาะสมและถูกต้องแม่นยำ และหาจำนวนกลุ่มประชากรย่อยที่เหมาะสมได้ โดยไม่คำนึงถึงสมติฐานใด ๆ ของจุดกำเนิดของกลุ่มประชากรย่อย ระเบียบวิธีใหม่นี้ได้ถูกทดสอบกับข้อมูลจำลองที่รู้กลุ่มประชากรที่แน่นอนแล้ว และยังได้ทดสอบ

กับข้อมูลจริงซึ่งได้ผลเป็นที่น่าพอใจ

สำหรับงานวิจัยในอนาคตที่เป็นไปได้คือการพัฒนาระเบียบวิธีนี้ให้สามารถหาอัตตราส่วนการผสมกันของกลุ่มประชากรย่อยได้ ซึ่งขณะนี้ทางคณะผู้ทำการวิจัยได้ทำการพัฒนาอยู่ ซึ่งถ้าทำสำเร็จจะเป็นการใช้ PCA สำหรับการแก้ปัญหาทางประชากรได้อย่างมีประสิทธิภาพสูงในระดับหนึ่ง

เอกสารอ้างอิง

1. Cardon LR, Palmer LJ: **Population stratification and spurious allelic association**. *Lancet* 2003, **361**(9357):598-604.
2. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data**. *Genetics* 2000, **155**(2):945-959.
3. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis**. *PLoS genetics* 2006, **2**(12):e190.
4. Guojun Gan CM, Jianhong Wu: **Data Clustering: Theory, Algorithms, and Applications**: SIAM, Society for Industrial and Applied Mathematics 2007.
5. Golub G, H., Van Loan F, C: **matrix computations**, 3 edn. Baltimore: The Johns Hopkins University Press; 1996.
6. Bezdec JC: **Pattern Recognition with Fuzzy Objective Function Algorithms**. New York: Plenum Press; 1981.
7. Liang L, Zollner S, Abecasis GR: **GENOME: a rapid coalescent-based whole genome simulator**. *Bioinformatics (Oxford, England)* 2007, **23**(12):1565-1567.
8. Gao X, Starmer JD: **AWclust: point-and-click software for non-parametric population structure analysis**. *BMC bioinformatics* 2008, **9**:77.
9. **http://hapmap.org**
10. **ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp/Btau20040927**
11. Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, Schurr TG, Zhadanov SI, Osipova LP, Brutsaert TD *et al*: **Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation**. *Human genomics* 2005, **2**(2):81-89.
12. **http://www.ansi.okstate.edu/breeds/cattle/**
13. Reich D, Price AL, Patterson N: **Principal component analysis of genetic data**. *Nature genetics* 2008, **40**(5):491-492.
14. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, Drineas P: **PCA-correlated SNPs for structure identification in worldwide human populations**. *PLoS genetics* 2007, **3**(9):1672-1686.
15. Falush D, Stephens M, Pritchard JK: **Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies**. *Genetics* 2003, **164**(4):1567-1587.
16. Pritchard JK, Rosenberg NA: **Use of unlinked genetic markers to detect population stratification in association studies**. *American journal of human genetics* 1999, **65**(1):220-228.
17. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL *et al*: **Worldwide human relationships inferred from genome-wide patterns of variation**. *Science (New York, NY* 2008, **319**(5866):1100-1104.
18. Lee C, Abdool A, Huang CH: **PCA-based population structure inference with generic clustering algorithms**. *BMC bioinformatics* 2009, **10 Suppl 1**:S73.
19. Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ: **Recent genetic selection in the ancestral admixture of Puerto Ricans**. *American journal of human genetics* 2007, **81**(3):626-633.

ภาคผนวก

Manuscript Submitted to BMC Bioinformatics

Iterative Pruning PCA Improves Resolution of Highly Structured Populations

Apichart Intarapanich[1], Philip J. Shaw[2], Anunchai Assawamakin[3,2], Pongsakorn Wangkumhang[2], Chumpol Ngamphiw[2], Kridsadakorn Chaichumpoo[2], Jittima Piriyapongsa[2], Sissades Tongsima[2§]

[1] NECTEC 112 Thailand Science Park, Paholyothin Road, Klong 1, Klong Luang, Pathumtani 12120 Thailand

[2] BIOTEC 113 Thailand Science Park, Paholyothin Road, Klong 1, Klong Luang, Pathumtani 12120 Thailand

[3] Division of Molecular Genetics, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Bangkok 10700 Thailand

[§]Corresponding author

Email addresses:

AI: apichart.intarapanich@nectec.or.th

PJS: philip@biotec.or.th

AA: anunchai_ice@yahoo.com

PW: pongsakorn.wan@biotec.or.th

CN: chumpol.nga@biotec.or.th

KC: kridsadakorn.cha@biotec.or.th

JP: jittima.pir@biotec.or.th

ST: sissades@biotec.or.th

Abstract

Background

Non-random patterns of genetic variation exist among individuals in a population owing to a variety of evolutionary factors. Therefore, populations are structured into genetically distinct subpopulations. As genotypic datasets become ever larger, it is increasingly difficult to correctly estimate the number of subpopulations and assigning individuals to them. The computationally efficient non-parametric, chiefly Principal Components Analysis (PCA)-based methods are thus becoming increasingly relied upon for population structure analysis. Current PCA-based methods can accurately detect structure; however, the accuracy in resolving subpopulations and assigning individuals to them is wanting. When subpopulations are closely related to one another, they overlap in PCA space and appear as a conglomerate. This problem is exacerbated when some subpopulations in the dataset are genetically far removed from others. We propose a novel PCA-based framework which addresses this shortcoming.

Results

A novel population clustering program called iterative pruning PCA (ipPCA) was developed which assigns individuals to subpopulations and infers the total number of subpopulations present. Genotypic data from simulated and real population datasets with different degrees of structure were analyzed by the proposed method. For datasets with simple structures, the subpopulation assignments of individuals made by ipPCA were consistent with the STRUCTURE and AWclust algorithms. On the other hand, highly structured populations containing many closely related subpopulations could be accurately resolved only by ipPCA, and not by other methods.

Conclusions

The algorithm is computationally efficient and not constrained by the dataset complexity. This systematic clustering approach removes the need for prior population labels, which could be advantageous when cryptic stratification is encountered in datasets of individuals otherwise assumed to be homogenous.

Background

Allele frequencies vary across populations because of systematic differences in ancestry; these differences arise from many factors such as migration, selection and drift. Hence, populations are genetically substructured. The information obtained from resolving population substructure can be used to infer population history. Furthermore, human disease association studies must account and correct for the population substructure to reduce spurious associations and reveal the predisposing factors of disease [1]. Analysis of population stratification must meet four main challenges namely: (i) detecting structure, (ii) assigning individuals to subpopulations, (iii) determining the number of optimal, or primal, subpopulations (K) and (iv) determining the extent of subpopulation admixture [2].

With the advent of high throughput genotyping, increasingly large genotypic datasets (e.g. HapMap dataset of 3.5 million single nucleotide polymorphism (SNP) arrays from 270 individuals [3]) will provide progressively difficult challenges for population structure analysis. Therefore, to keep abreast with the ever increasing size and complexity of genotypic data, refinement of existing analytical methods and entirely novel approaches will be needed to resolve subpopulations. Several different methods have been proposed to address some aspects of the population substructure problem. These methods can be categorized into two main approaches, namely parametric and non-parametric-based. STRUCTURE is the “gold standard” parametric-based algorithm for population stratification analysis [2, 4], because it can address all four substructure analysis challenges. However, STRUCTURE imposes high computational burden and it is thus impractical for the large datasets typically analyzed nowadays. STRUCTURE also has other weaknesses, mainly its inference of

primal *K*, which requires extensive statistical testing of several STRUCTURE runs performed with increasing *K*. STRUCTURE's inference of *K* can also vary according to the model used [4], hence the determination of admixture, which is highly dependent on the *K* value used needs to be interpreted with caution. Other parametric methods, such as L-POP [5], PSMIX [6] and *frappe* [7] employ different parametric approaches and are more computationally efficient, although they address only specific population structure problems not adequately solved by STRUCTURE, such as admixture [5].

For non-parametric methods, the algorithms in this class do not require model assumptions. Principal Components Analysis (PCA) is the most widely used method for visualizing structure which uses a covariance matrix for eigenanalysis, allowing representation of individuals as data points in scatter-plots. Principal Coordinate Analysis (PCoA) is an alternative method for eigenanalysis which uses an allele-sharing distance (ASD) matrix, and gives different scatter-plot patterns from PCA [8, 9]. PCA is not a method for assigning individuals nor estimating *K* and is often used merely to visualize the population structure trend. The most popular PCA-based algorithm applied to population structure analysis is EIGENSTRAT/SmartPCA [10, 11], which has been used by several investigators for large datasets typically required for studies of human population structure and genome association [12-14]. This algorithm employs a computationally-efficient variant of eigenanalysis to report the probability of population substructure according to Tracy-Widom (TW) distribution.

In a typical population dataset, genetic distances vary among subpopulations and PCA scatter-plots can reveal the most genetically isolated subpopulations as distinct clusters of individuals in a small number of principal components. Hence, supervised clustered with prior assumption of the number of *K* subpopulation clusters

can be done to assign individuals. Conversely, closely related subpopulations will occupy a confined feature space and appear as a conglomerate. For example, in the scatter plot of PC1 versus PC2, conglomerates containing individuals with different population labels are frequently observed. In some cases, the distinction between closely related subpopulations is apparent in a greater number of principal components [15]. Thus, in order to resolve closely related subpopulations, individuals must be separated using a clustering algorithm working in multidimensional PCA space [16-18]. However, clustering algorithms perform inconsistently when too many axes of variation are used [19]. Clustering algorithms require separation in axes of variation. However, clusters in some axes may merge into a single cluster, and hence clustering algorithms can become confused. Generally the informative axes of variation are contained within the rank of matrix [20]. Therefore, the number of principal components should be optimized and not exceed a certain number for each dataset.

One way to improve the resolution of closely-related subpopulations in PCA scatter-plots is by removing genetically distant individuals from the dataset. In the investigation of European human genetic substructure using 300K SNP arrays [21], it was found that prior exclusion of individuals belonging to certain groups improved substructure resolution of the other groups, e.g. removal of Ashkenazi Jewish individuals led to clearer substructuring of other northern European groups. This data "cleaning" approach is clearly advantageous, although the method as described in [21] is *ad hoc* and unable to detect and remove subtle outliers [22]. Clearly, this approach is not feasible for datasets composed of individuals presumed to belong to a homogenous population without any distinguishing labels, for instance disease association studies carefully controlled for ethnicity and geographical origin. With

sufficient number of markers and individuals, cryptic structure in an apparently homogenous population can be detected using PCA [10, 11], which cannot be resolved by current unsupervised clustering methods, with no assumptions of *K* [18].

To determine the primal *K* for unsupervised clustering, the gap statistic [23] is employed on the AWclust results [16] and Density-Based Means clustering results [18]. Alternatively, the Bayesian Information Criterion (BIC) approach for determining *K* can be applied to the clustering results [17]. Calculating the gap statistic and BIC are computationally intensive, thus these approaches are impractical for highly structured datasets. Furthermore, these approaches are sensitive to noisy non-informative PCs, which may explain why they are not appropriate for highly structured datasets with a large *K*. All unsupervised clustering approaches currently applied to PCA do not assign individuals into subpopulations according to genetic distance between subpopulations in a fully comprehensive and systematic fashion. These algorithms thus have insufficient discriminatory power to assign individuals to subpopulations and determine *K* for highly structured datasets.

In this study, we propose a non-parametric analytical framework which incorporates several key refinements of the PCA-based approach. The new algorithm, which we call iterative pruning PCA (ipPCA) addresses three of the main challenges for population structure analysis, namely detection of structure, assigning individuals to subpopulations and determining the primal *K* with greater accuracy than previously proposed algorithms. The ipPCA algorithm utilizes a novel unsupervised clustering heuristic which markedly improves resolution of population substructure by an iterative process. In this method, individuals are systematically bisected according to genetic similarity at each step, continuing until all subpopulations are revealed.

Results

Algorithm

The ipPCA algorithm is a PCA framework which utilizes non-redundant principal components to construct a transformed domain of the input data, which can be mapped to PCA space. By means of selecting a limited number of principal components, dimension reduction can be achieved. The PCA domain allows each input individual to be represented as a single datum point in a scatter-plot, or in clustering analysis. The ipPCA technique constructs the transformed domain based on a covariance matrix of the data matrix containing SNP genotypic data encoded as 0, 1 or 2 elements.

$$X = \overbrace{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NM} \end{pmatrix}}^{\text{GenotypeData}} = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_N \end{pmatrix}$$

where *M* and *N* are the number of markers and individuals respectively. Note that *M* is normally much larger than *N*.

Since the covariance matrix is usually a large square matrix, it is memory and computationally intensive to compute the PCA space. However, there is an alternative technique to compute the transformed domain without computational burden [20]. Instead of using the data covariance matrix, the PCA space can be constructed by decomposing a modified data matrix XX^T, which is much smaller than the covariance matrix. For example, a dataset with 200 individuals and 1000000 markers would have 1000000×1000000 elements in the covariance matrix compared to 200×200 elements in the XX^T matrix. The modified data matrix can be factored as $USV^T = XX^T$. Then, the eigenvectors can be computed by

$$E = S(1:k, 1:k)^{-1} U(:, 1:k) X$$

where k is the rank of the data matrix and $E = [\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_k]$ are the eigenvectors. Here, we use the "colon" notation [20] to specify a column or row of a matrix. Then, the PCA space is constructed based on these eigenvectors.

Because ipPCA is sensitive to the data pattern, data quality is crucial for accuracy of the algorithm. Typically, it is necessary to clean the data before performing ipPCA. Technical limitations in genotyping occasionally lead to missing values at some loci. To check whether missing data by itself can create substructure, all missing data are encoded as zero and other loci are encoded as one. The zero and one (abbreviated as 0-1) encoded dataset is analyzed by SmartPCA [10]. If substructure can be observed as clear outlier individuals on the periphery of the two PC scatter-plot, these individuals can be removed from the dataset, as suggested by [10].

Before performing ipPCA, a quality control check is performed on the 0-1-2 encoded data matrix. In this case, frequency counts are computed on each locus to ensure that entries encoded as zero are homozygous major allele (wild-type), entries encoded as one are heterozygous and entries encoded as two are homozygous minor allele. The pre-processing steps and the ipPCA framework can be summarized as follows:

Pre-processing steps:

Check if missing values in the input dataset cause detectable substructure using EIGENSTRAT/SmartPCA and if significant substructure is reported, remove individuals with missing data which cause substructure (identified as outliers on PC1-PC2 scatter-plot).

ipPCA steps:

1. Make matrix X_i for each *non-terminating* data group

2. Use singular value decomposition (SVD) [20] to factorize the $X_i X_i^T$ data into $U_i S_i V_i^T$

3. Project all individuals into PC space using the number of PCs equal to the data matrix rank [15].

4. Check terminating condition on the S_i using the TW test statistic implemented in EIGENSTRAT/SmartPCA ([10])

   (a) Terminate if TW test statistic is insignificant for the first PC ($p>10^{-12}$)

   (b) Otherwise, proceed to the next step

5. Apply fuzzy *c*-means [24] to split individuals X_i into two clusters

6. Repeat from step 1 until all the data are terminated

7. When all the iterative processes are terminated, the number of subpopulations can be determined by counting all the terminal nodes (determination of *K*). Note that replicated ipPCA runs are typically performed to test the robustness of the clustering algorithm.

Testing

The power and robustness of the proposed ipPCA algorithm was explored and optimized using simulated datasets. The performance of ipPCA was then tested on three real datasets, namely HapMap, bovine, and Shriver's datasets. Finally, we compared the results of our algorithm with the results from existing tools: STRUCTURE and AWclust. STRUCTURE version 2.2 was downloaded from J. Pritchard's website [25] and applied using the following parameters: 100000 burn-ins, 100000 runs, with admixture model, no LD model. For AWClust analysis, AWClust was downloaded from [26] and the algorithm's default parameters were selected.

Iterative pruning PCA clustering: a way to improve discriminatory power

The program GENOME [27] was employed to generate simulated genotypic data under the Wright-Fisher neutral coalescent model (backward in time) [28]. Three evolutionary models (called population histories in GENOME) were constructed (Figure 1). The model 1 dataset (Figure 1A) contains three subpopulations derived from two ancestral populations. The model 2 dataset (Figure 1B) contains five independent subpopulations with no admixture. The model 3 datasets (Figure 1C) contain 20 subpopulations derived from three ancestral populations. Parameter settings used to generate each model are as follows:

The model 1 dataset parameters:

```
-pop 3 50 50 50 –c 20 –s 500 –N model1.txt
```

The model 2 dataset parameters:

```
-pop 5 50 50 50 50 50 –c 20 –s 500 –N model2.txt
```

The model 3 dataset parameters:

```
-pop 20 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50
-c 20 –s 500 –N model3.txt
```

Using the above parameters, we generated 10000 SNPs for each simulated dataset. Note that the model 3 population history was used to generate 30 datasets to test the robustness of ipPCA. All simulated datasets and the population history files (model1.txt, model2.txt, and model3.txt) analyzed in this study can be downloaded from [29].

The performance and robustness of ipPCA to resolve structure were investigated using simulated datasets of increasing complexity. Two simple population structure models (K=3 and 5 subpopulations) and one complex model

(*K*=20 subpopulations) were simulated using the GENOME tool [27] . For the first two simple models (models 1 and 2), ipPCA can resolve *K* consistent with the models and assign all individuals correctly to the corresponding subpopulations (Figures 2 and 3). For the highly complex model (model 3), 30 simulated datasets were generated using the same evolutionary model. The structure resolved by ipPCA was highly consistent with the model, for both number of inferred *K* and the individuals assigned to each subpopulation (mode *K*=20, range 19-22; mean mis-assignment rate 6.07%, range 1.9%-17.5%), see additional file 1). To test the reproducibility of the clustering algorithm, ten replicated ipPCA experiments were performed on each data set. The results were reproduced exactly for all ten runs in each model (data not shown).

Assignment of individuals to subpopulations in real datasets is reproducible and consistent with population labels

Three real datasets, namely HapMap, bovine, and Shriver's, were analyzed in this work. The HapMap dataset, retrieved from [30], comprises individuals with four population labels: Han Chinese from Beijing (CHB), Japanese from Tokyo (JPT), Caucasian European from Utah (CEU), and Yoruba from Ibadan (YRI) with 1533661 SNPs. Fifty thousand SNPs were uniformly re-sampled from this SNP pool [3]. Starting from the first marker, a moving window was used to select SNPs in an even spacing fashion, such that every 30^{th} marker was selected from the 1533661 markers (for full list of SNP markers used see additional file 2). Data pre-processing ipPCA was performed and two outlier JPT individuals were removed.

The bovine dataset was downloaded from [31]. The bovine SNP data (9329 SNPs) are publicly available as part of the Bovine Genome Project [32]. After data pre-

processing, no outliers were detected. Shriver's worldwide human SNP dataset of 307 individuals with 14 different ethnic/geographical labels was provided by Prof. Mark D. Shriver, which is a dataset expanded from the one originally published in [33]. This dataset consists of 11555 SNPs for each individual, evenly distributed over the entire genome. After data pre-processing, no outliers were detected.

The HapMap dataset was analyzed by ipPCA, which contains individuals from four distinct geographical regions (YRI, CEU, CHB, and JPT). All individuals in each of the four subpopulations defined by ipPCA share the same distinguishing population label (Figure 4). The YRI subpopulation was defined in iteration 1 while the CEU was defined in iteration 2. Finally, the CHB and JPT subpopulations were defined in iteration 3. Exactly the same assignment results were obtained from for ten replicated runs using 50000 SNPs and a single ipPCA run using the entire 1533661 SNP collection (data not shown).

The bovine dataset contains individuals from nine breeds considered to be genetically distinct subpopulations. The Brahman (BRM) breed cattle are zebu (also known as *B. indicus*), considered a sub-species very distinct from the taurine cattle breeds. Santa Gertrudis (SGT) is a composite breed derived from BRM and Shorthorn breed cattle. The other breeds, Angus (ANG), Charolais (CHL), Limousin (LMS), Hereford (HFD), Norwegian red (NRC), Jersey (JER) and Holstein (HOL) are taurine cattle and European in origin (for breed descriptions see [34]).

Scatter-plot analysis of the entire bovine dataset showed that individuals from the HFD, JER, HOL, LMS, ANG, CHL and NRC taurine breeds appear as a conglomerate and are separated from BRM and SGT individuals (Figure 5). Figure 6 shows the resulting consensus subpopulation tree from ten replicated ipPCA runs, which all gave the same results (data not shown). Each ipPCA run requires six

successive iterations of ipPCA to define all subpopulations. After the second iteration, the BRM and SGT cluster was divided satisfying the termination criterion, thus defining two subpopulations composed of BRM and SGT individuals, respectively. Additionally, a third subpopulation composed of JER individuals was defined and separated from the taurine cluster. After more iterations of ipPCA, a further six subpopulations were defined in which each subpopulation is largely composed of individuals of the same breed.

Shriver's dataset was then analyzed by ipPCA. From the scatter-plot analysis of the entire dataset, it can be observed that individuals are broadly grouped according to their geographical origins with some overlap between individuals with different labels (Figure 7). The consensus ipPCA result showed that after the second iteration of ipPCA, a subpopulation composed of African Americans and a Puerto Rican individual was defined. After further iterations of ipPCA, more subpopulations were defined with each containing individuals with shared geographical and/or ethnic origins. After six iterations, twelve subpopulations were defined with no further substructure found (see Figure 8). Subpopulation assignments were robust. Only one ipPCA run differed from the consensus (see additional file 3)

The number of predicted *K* and assignment of individuals to subpopulations were compared with other algorithms. Table 1 shows comparison between different algorithms for inferring *K* from 35 datasets. All algorithms agreed with the same *K* value only for the least complex dataset (model 1). For all others, there was no concordance between the algorithms. Given the discordances between the algorithms we then investigated the individual assignments made by each algorithm.

For datasets of low complexity, assignment of individuals to subpopulations was largely consistent among the STRUCTURE, AWclust, and ipPCA algorithms

(see additional file 4 for AWclust results and additional file 5 for STRUCTURE results). Given the greater discordance between algorithms for inference of *K* from highly structured datasets, a more detailed comparison of individual assignments to subpopulations was made between STRUCTURE and ipPCA. The comparison was made using Shriver's highly structured dataset for the optimal number of *K* reported by each algorithm (Figure 9). Each of the twelve subpopulations defined by ipPCA contain similar number of individuals (range 19-46), and the assignments of individuals to subpopulations SP-4, SP-5, SP-6, SP-7, SP-8, SP-9, SP-11 and SP-12 are consistent with the population labels. Four subpopulations, namely SP-1, SP-2, SP-3 and SP-10 contain individuals with more than one population label.

The 14 subpopulations assigned by STRUCTURE differed markedly to ipPCA in the number of individuals contained within each. Two subpopulations have no assigned individuals, five subpopulations have fewer than ten individuals and two subpopulations have more than 50 assigned individuals. Inspection of the population labels of the assigned individuals shows that STRUCTURE assigned a subpopulation containing individuals from three geographically disparate African locations together with African Americans and a Puerto Rican in contrast to ipPCA, which assigns African Pygmy, West African and East African individuals as separate subpopulations SP-4, SP-8 and SP-9, respectively. STRUCTURE also assigned Caucasian, Puerto Rican, South Altaian and West African individuals among more subpopulations than ipPCA. STRUCTURE also grouped Nahua and Quechua/Peru individuals as one subpopulation and Asian and South Altaian individuals as another subpopulation. However, ipPCA assigned these individuals to four separate subpopulations consistent with population labels (SP-5, SP-6, SP-11, SP-12).

In order to demonstrate the power of the iterative clustering approach, a comparison was made between the individual subpopulation assignments by ipPCA and PCA clustering done in a *non-iterative* fashion (third column in Figure 9). The PCA results from Shriver's entire dataset were clustered using the same fuzzy *c*-means clustering algorithm to assign individuals into 12 clusters (the number inferred by ipPCA). The number of clusters to be used for clustering was not calculated independently, e.g. using the gap statistic since this was not feasible for such a complex dataset. It can be observed that clustering on non-iterative PCA (NI-PCA) results leads to notable differences in individual assignment. For instance, there are three groups with fewer than ten individuals (3 Puerto Ricans, 8 East Africans, and 4 New Guinea) and one conglomerate group containing 68 individuals (4 Spanish, 11 South Indian Mala, 11 South Indian Brahman, and 42 Caucasians). Furthermore, Asian and South Altaian individuals were assigned together, whereas East African and New Guinea individuals were each separated into two subpopulations.

Implementation

ipPCA was implemented in MATLAB version 2007a. It is available in both graphical user interface (GUI) and MATLAB function. GUI ipPCA is suitable for small datasets, e.g. 50000 makers, while MATLAB function ipPCA can be used for larger datasets. The source codes for both versions are available on request.

Discussion

Practicality of ipPCA comparing with other algorithms

We have demonstrated a novel PCA-based analytical framework that accurately resolves population stratification including that of highly structured datasets. With this tool, individuals are assigned to subpopulations and *K* is determined with high accuracy. Moreover, minimal computational effort is required. The computational time and the complexity of the datasets were not limitations for ipPCA, unlike other algorithms. For example, AWclust limits the number of inferred *K* to seven or less while STRUCTURE takes too long to compute when *K* and the number of markers are large. For datasets in which comparison can be made between different algorithms, the *K* values reported by ipPCA were consistent with other algorithms (STRUCTURE and AWclust) and the presumed *K*. For highly structured datasets requiring many SNPs, the proposed algorithm performs faster than STRUCTURE and AWclust algorithms. For example, Shriver's dataset requires approximately three days to compute by STRUCTURE while ipPCA needed less than ten minutes on a 32-core AMD 2.3 gigahertz Opteron with 64 gigabytes of memory running Linux CentOS operating system.

Improved clustering through PC selection

The trend observed among datasets analyzed in this study is that early iterations require more PCs for clustering than later ones (Figures 2, 3, 4, 6, and 8). This is in agreement with the findings of the authors of EIGENSTRAT/SmartPCA, who showed that the number of significant eigenvectors (PCs) to resolve structure reflects the number of subpopulations [10, 35]. However, the number of *K* is not simply defined by the number of significant PCs, since the actual number of PCs needed to reveal subpopulations varies in each case. In ipPCA, the clustering process utilizes the optimal number of PCs (determined to be the matrix rank), which varies among

nested datasets according to the number of individuals at each iteration. After each iteration of ipPCA, the nested datasets have progressively simpler structures so that fewer principal components are used for clustering.

Inference of K

The inference of the number of optimal, or primal, *K* subpopulations is critically dependent on the assignment accuracy of individuals. High assignment accuracy allows the correct value of primal *K* to be inferred. The STRUCTURE and AWclust algorithms have lower assignment accuracy than ipPCA, and thus their inferences of *K* were incorrect. For example, STRUCTURE and AWclust reported $K=3$ for the HapMap dataset and could not resolve CHB and JPT as separate subpopulations. The resolution of CHB and JPT subpopulations by ipPCA was consistent with the finding of [36], who identified the subset of informative markers for separating these individuals as two subpopulations.

In order to define subpopulations from the PCA clustering results, ipPCA uses a novel approach which considers only the eigenvalue distribution. By this approach, subpopulations are defined by a standard criterion thus removing subjectivity from the definitions. Rigorous subpopulation definitions are currently lacking, since there is no agreement as to precisely what genetic variation accounts for population structure [37]. The difficulty in defining subpopulations is the main reason why determining *K* is considered very challenging [2, 4].

Assignment accuracy

In terms of individual assignments, there were some notable differences between ipPCA and STRUCTURE for Shriver's dataset. For the optimal number of

subpopulations reported by STRUCTURE (K=14) (see additional file 6), assignments were inconsistent with what have been reported by other methods [33, 38]. From a population evolutionary perspective, some of the assignments made by STRUCTURE were not meaningful. For instance, there were two groups that STRUCTURE assigned no individuals to, while five other groups contained fewer than ten individuals (Figure 8). By limiting K to twelve as what ipPCA predicted, the assignments made by STRUCTURE changed with removal of the empty groups and amalgamation of south Altaian individuals into one subpopulation (see additional file 7). However, some inconsistencies with accepted patterns of human population structure remained. For example, the subpopulation with pan-African individuals remained.

The ipPCA clustering approach differs from the approaches used by others for clustering PCA results, e.g.[17] who performed clustering on the entire dataset using a large number of PCs. For highly structured populations, this kind of approach is not able to accurately assign individuals to subpopulations, irrespective of the clustering algorithm used (e.g. *k*-means, soft *k*-means, spectral *k*-means) since the closely related subpopulations are confined in a small region of feature space (see Background section). The greatest problem for the NI-PCA approach is in knowing the number of clusters to be used in clustering. Indeed, the analysis of highly structured datasets in this paper shows that NI-PCA approach fails to accurately assign individuals, even when using the correct K inferred by ipPCA (for analysis of simulated data Model 3, see additional file 8). We chose fuzzy *c*-means for clustering since the number of clusters in each ipPCA iteration is restricted to two, hence a simple algorithm is sufficient. Furthermore, the cluster centroids determined by fuzzy *c*-means are more

consistent compared with the commonly used *k*-means algorithm [39], which is most important for subpopulation assignment in our case.

Definition of subpopulations according to genetic distance and population history

The iterative pruning approach also defines subpopulations in a manner reflecting the probable evolutionary origin of each subpopulation. As can be seen from the analysis of simulated datasets (Figure 2 and additional file 1) and real datasets in Figures 4-8, closely related subpopulations are defined after more distantly related ones. This iterative pruning process thus offers a systematic way to define subpopulations according to their degrees of relatedness to one another. By removing the most distant individuals to create nested datasets, one is able to resolve substructure, which would not be revealed otherwise.

For more complex datasets containing subpopulations with individuals of recent admixed origins, these were shown by ipPCA as those which contain assigned individuals with different population labels. For instance, two subpopulations were defined from the bovine dataset containing individuals with different breed labels. The overlap of NRC and HOL individuals in a subpopulation is not unexpected, since NRC is not a pure breed and has HOL ancestry. Similarly, the assignment of some SGT individuals to a subpopulation with taurine cattle reflects the fact that SGT is a composite zebu/taurine breed.

For Shriver's dataset which has a more complex structure than the bovine dataset, twelve subpopulations were defined by ipPCA with individuals assigned consistent with the group labels and the neighbor-joining tree shown in [33]. Four

subpopulations appeared to be admixed containing individuals with different labels. Among these four admixed subpopulations, Puerto Rican individuals were assigned to two subpopulations. These assignments reflect the Puerto Rican population history, in which their genomes still contain the signatures of ancestral European and African migrants [33, 40]. We are currently developing an admixture extension to ipPCA, which is outside the scope of this paper. In this paper, we wished to provide a solid platform for ipPCA to demonstrate its accuracy in resolving all *K* primal subpopulations, which is crucial for admixture testing.

Conclusions

We propose some key refinements to the PCA-based algorithm for improved resolution of population genetic stratification. With this new method, individuals can be accurately assigned to subpopulations systematically, thus defining the optimal, or primal, *K* without any assumptions of individuals' origins or degree of relatedness to one another. The power of the technique was demonstrated using datasets with known structure, although we think there is potential to reveal structure in datasets with cryptic stratification.

Authors' contributions

AI, ST wrote algorithms. AI, PJS, AA, PW, CN, KC, JP, ST analyzed data. AI, PJS, ST wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Prof. Mark D. Shriver for providing access to the worldwide human SNP dataset presented in this work. We also thank Dr. Liming Liang for helping us with the GENOME software. We acknowledge the Cluster and Program Management Office, National Science and Technology Development Agency (Grant No. BT-B-02-IM-GI-5101). AI was supported by Thailand Research Fund (Grant No. TRG5180005). AA was partially supported by the Royal Golden Jubilee Ph.D. Program (Grant No.PHD/4.I.MU.45/C.1). Finally, PJS, PW, CN, KC, JP, and ST were supported by BIOTEC.

References

1. Cardon LR, Palmer LJ: **Population stratification and spurious allelic association**. *Lancet* 2003, **361**(9357):598-604.
2. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data**. *Genetics* 2000, **155**(2):945-959.
3. Consortium IH: **A haplotype map of the human genome**. *Nature* 2005, **437**(7063):1299-1320.
4. Falush D, Stephens M, Pritchard JK: **Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies**. *Genetics* 2003, **164**(4):1567-1587.
5. Purcell S, Sham P: **Properties of structured association approaches to detecting population stratification**. *Human heredity* 2004, **58**(2):93-107.
6. Wu B, Liu N, Zhao H: **PSMIX: an R package for population structure inference via maximum likelihood method**. *BMC bioinformatics* 2006, **7**:317.
7. Tang H, Peng J, Wang P, Risch NJ: **Estimation of individual admixture: analytical and study design considerations**. *Genetic epidemiology* 2005, **28**(4):289-301.
8. Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD: **Measuring European population stratification with microarray genotype data**. *American journal of human genetics* 2007, **80**(5):948-956.
9. Reeves PA, Richards CM: **Accurate Inference of Subtle Population Structure (and Other Genetic Discontinuities) Using Principal Coordinates**. *PLoS ONE* 2009, **4**(1).
10. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis**. *PLoS genetics* 2006, **2**(12):e190.

11. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nature genetics* 2006, **38**(8):904-909.
12. Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, Hankinson SE, Hu FB, Duffy DL, Zhao ZZ *et al*: **A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation**. *PLoS genetics* 2008, **4**(5):e1000074.
13. Liu Y, Helms C, Liao W, Zaba LC, Duan S, Gardner J, Wise C, Miner A, Malloy MJ, Pullinger CR *et al*: **A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci**. *PLoS genetics* 2008, **4**(3):e1000041.
14. Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C, Filsell W, Ginger RS, Green MR, van der Ouderaa FJ *et al*: **A genomewide association study of skin pigmentation in a South Asian population**. *American journal of human genetics* 2007, **81**(6):1119-1132.
15. Parsons L, Haque E, Liu H: **Subspace Clustering for high dimensional data : A review**. *Sigkdd Explorations* 2004, **6**(1):15.
16. Gao X, Starmer JD: **AWclust: point-and-click software for non-parametric population structure analysis**. *BMC bioinformatics* 2008, **9**:77.
17. Lee C, Abdool A, Huang CH: **PCA-based population structure inference with generic clustering algorithms**. *BMC bioinformatics* 2009, **10 Suppl 1**:S73.
18. Liu N, Zhao H: **A non-parametric approach to population structure inference using multilocus genotypes**. *Human genomics* 2006, **2**(6):353-364.
19. Agrawal R, Gehrke J, Gunopulos D, Raghavan P: **Automatic Subspace Clustering of High Dimensional Data for data mining applications**. *SIGMOD Record ACM Special Interest Group on Management of Data* 1998, **27**(2):94-105.
20. Golub G, H., Van Loan F, C: **matrix computations**, 3 edn. Baltimore: The Johns Hopkins University Press; 1996.
21. Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi L, Gregersen PK *et al*: **Analysis and application of European genetic substructure using 300 K SNP information**. *PLoS genetics* 2008, **4**(1):e4.
22. Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A *et al*: **On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants**. *American journal of human genetics* 2008, **82**(2):453-463.
23. Tibshirani R WG, Hastie T: **Estimating the number of clusters in a dataset via the gap statistic**. *Journal Royal Statistical Soc B* 2001, **63**:411-423.
24. Bezdec JC: **Pattern Recognition with Fuzzy Objective Function Algorithms**. New York: Plenum Press; 1981.
25. **http://pritch.bsd.uchicago.edu/software/structure2_2.html**
26. **http://awclust.sourceforge.net/**
27. Liang L, Zollner S, Abecasis GR: **GENOME: a rapid coalescent-based whole genome simulator**. *Bioinformatics (Oxford, England)* 2007, **23**(12):1565-1567.
28. Ewens WJ: **Mathematical Population Genetics.** Berlin: Springer; 1979.
29. **http://www.biotec.or.th/gi/tools/ippca**
30. **http://hapmap.org**

31. **ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp/Btau20040927**
32. **http://www.hgsc.bcm.tmc.edu/projects/bovine/index.html**
33. Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, Schurr TG, Zhadanov SI, Osipova LP, Brutsaert TD *et al*: **Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation**. *Human genomics* 2005, **2**(2):81-89.
34. **http://www.ansi.okstate.edu/breeds/cattle/**
35. Reich D, Price AL, Patterson N: **Principal component analysis of genetic data**. *Nature genetics* 2008, **40**(5):491-492.
36. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, Drineas P: **PCA-correlated SNPs for structure identification in worldwide human populations**. *PLoS genetics* 2007, **3**(9):1672-1686.
37. Waples RS, Gaggiotti O: **What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity**. *Molecular ecology* 2006, **15**(6):1419-1439.
38. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL *et al*: **Worldwide human relationships inferred from genome-wide patterns of variation**. *Science (New York, NY* 2008, **319**(5866):1100-1104.
39. Guojun Gan CM, Jianhong Wu: **Data Clustering: Theory, Algorithms, and Applications**: SIAM, Society for Industrial and Applied Mathematics 2007.
40. Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ: **Recent genetic selection in the ancestral admixture of Puerto Ricans**. *American journal of human genetics* 2007, **81**(3):626-633.

Figures

Figure 1. Population history trees for generating simulated datasets.

The GENOME tool [27] was used to generate the simulated datasets.

A) three subpopulations mixed model (model 1)

B) five independent subpopulations (model 2)

C) twenty subpopulations (model 3)

Figure 2. ipPCA analysis of simulated data model 1 with 3 subpopulations.

A) Consensus subpopulation tree from ten replicated ipPCA runs. Each cell contains labels SP-1, SP-2 or SP-3 referring to subpopulation labels used in simulation. The number of individuals is presented in the parentheses next to each label. The number of PCs used for clustering is indicated in the parentheses in each cell. The blue cell indicates the entire dataset. Non-terminated sub-datasets are presented in green cells while the terminated ones (resolved subpopulations) are in red cells.

B) Scatter-plot using the first and second principal components (PC1 vs. PC2) of the entire dataset (iteration 0 of ipPCA). Each datum point represents an individual. Each subpopulation label is denoted by a separate symbol (see inset).

Figure 3. ipPCA analysis of simulated data model 2 with 5 subpopulations.

A) Consensus subpopulation tree on ten replicated ipPCA runs. Each cell contains labels SP-1, SP-2, SP-3, SP-4 or SP-5 referring to subpopulation labels used in simulation. The number of individuals is presented in the parentheses next to each label. The number of PCs used for clustering is indicated in the parentheses in each cell. The blue cell indicates the entire dataset. Non-terminated sub-datasets are

presented in green cells while the terminated ones (resolved subpopulations) are in red cells.

B) Scatter-plot using the first and second principal components (PC1 vs. PC2) of the entire dataset (iteration 0 of ipPCA). Each datum point represents an individual. Each subpopulation label is denoted by a separate symbol (see inset).

Figure 4. ipPCA analysis of HapMap human dataset

A) **Consensus subpopulation tree**. Each cell contains labels YRI, CEU, CHB, or JPT referring to population labels. The number of individuals is presented in the parentheses next to each label. The number of PCs used for clustering is indicated in the parentheses in each cell. The blue cell indicates the pre-processed dataset. Non-terminated sub-datasets are presented in green cells while the terminated ones (resolved subpopulations) are in red cells.

B) Scatter-plots using the first and second principal components (PC1 vs. PC2). Each datum point represents an individual. Each subpopulation label is denoted by a separate symbol (see inset). The blue frame contains a scatter plot of ipPCA iteration 0. The green frame contains a scatter plot of the non-terminated sub-dataset at iteration 1. Scatter plots of resolved subpopulations are framed in red.

Figure 5. The PCA scatter-plot of the entire bovine dataset for the zeroth iteration of ipPCA.

The plot was made using the first two principal components (PC1 vs. PC2). Each datum point represents an individual. Each subpopulation label is denoted by a separate symbol (see inset).

Figure 6. Bovine consensus subpopulation tree on ten replicated ipPCA runs

Each cell contains labels CHL, SGT, JER, HOL, BRM, NRC, HFD, LMS, or ANG referring to population labels. The number of individuals is presented in the parentheses next to each label. The number of PCs used for clustering is indicated in the parentheses in each cell. The blue cell indicates the pre-processed dataset. Non-terminated sub-datasets are presented in green cells while the terminated ones (resolved subpopulations) are in red cells.

Figure 7. The PCA scatter-plot of the entire Shriver's dataset for the zeroth iteration of ipPCA.

The plot was made using the first two principal components (PC1 vs. PC2). Each datum point represents an individual. Each subpopulation label is denoted by a separate symbol (see inset).

Figure 8. Shriver's consensus subpopulation tree on ten replicated ipPCA runs

Each cell contains labels African American, African Pygmy, Asian, Caucasian, East African, Nahua, New Guinea, Puerto Rican, Quechua/Peru, South Altaian, South Indian Brahmin, South Indian Mala, Spanish, or West African referring to population labels. The number of individuals is presented in the parentheses next to each label. The number of PCs used for clustering is indicated in the parentheses in each cell. The

blue cell indicates the pre-processed dataset. Non-terminated sub-datasets are presented in green cells while the terminated ones (resolved subpopulations) are in red cells.

Figure 9. Population assigment comparison of Shriver's dataset among ipPCA, STRUCTURE, and non-iterative PCA clustering algorithms.

A) Log probability data plot of inferred *K* for STRUCTURE results with admixture model for *K*=8 to *K*=20; STRUCTURE parameters: 100000 simulation iterations for burn-ins and 100000 additional iterations for parameter estimation.

B) Side-by-side comparison of subpopulation assignment by ipPCA, STRUCTURE, and non-iterative PCA clustering (NI-PCA). Population labels given in [33] are abbreviated as follow: AFA (African American); AFP (African Pygmy); ASN (Asian); CCS (Caucasian); EAF (East African); NAH (Nahua); NGU (New Guinea); PTR (Puerto Rican); QPE (Quechua/Peru); SAT (South Altaian); SIB (South Indian Brahmin); SIM (South Indian Mala); SPN (Spanish); WAF (West African).

Tables

Table 1 - Comparison of three algorithms (AWclust, STRUCTURE, and ipPCA) for inferring the number of primal subpopulations (*K*).

Dataset	AWclust	STRUCTURE	ipPCA
model 1 simulated (3 subpopulations)	3	3	3
model 2 simulated (5 subpopulations)	5	6	5
model 3 simulated (20 subpopulations, 30 datasets)	N/A[1]	N/D[2]	20[3]
HapMap (4 population labels with 50000 SNPs)	3	3	4
bovine (9 population labels with 9329 SNPs)	N/A[1]	10	9
Shriver's (14 population labels with 11555 SNPs)	N/A[1]	14[4]	12

[1] Not applicable since AWclust gap statistics limits $K < 8$

[2] Not done owing to the STRUCTURE computational constraint

[3] Modal value from 30 simulated datasets

[4] There are two subpopulation groups with no individuals assigned to them

Additional files

Additional file 1
Title-Figure S1-S30

File format- compressed(rar file)-pdf

Description- Simulated data Model 3 ipPCA 30 runs results

Additional file 2
Title-Text S1

File format- compressed (rar file)-txt

Description- Full list of HapMap SNPs

Additional file 3
Title-Figure S31

File format-pdf

Description- Anomalous clustering result and subpopulation assignment from Shriver's dataset

Additional file 4
Title-Figure S32 – S34

File format- compressed (rar file)-pdf

Description- AWclust results of low complexity dataset

Additional file 5
Title-Text S2 – S5

File format-text-compressed (rar file)-text

Description- STRUCTURE results of low complexity dataset

Additional file 6
Title-Text S6

File format-text file

Description- STRUCTURE result of Shriver's data with $K = 14$

Additional file 7
Title-Figure S35

File format-pdf

Description- STRUCTURE result of Shriver's data with $K = 12$

Additional file 8
Title-Figure S36

File format-pdf

Description- Mis-assignment comparison between ipPCA and NI-PCA for Model 3 simulated datasets